STUDIES ON HIGH-SPEED CAPABILITY AND HIGH FUNCTIONALITY FOR CMOS LSI MEMORIES

by

Junichi Miyamoto

July 1988

STUDIES ON HIGH-SPEED CAPABILITY AND HIGH FUNCTIONALITY FOR CMOS LSI MEMORIES

by

Junichi Miyamoto

July 1988

Department of Electronics

Kyoto University

Table of Contents

1.Introduction 1

2.Design of a 256bit ECL RAM Using High-Speed Bipolar Technology 11

2.1. Overview 11

2.2. Cell Design and Process Sequence 13

2.3. Circuit Design 19

2.4. Summary 27

3.Design of a 64Kbit Bipolar-CMOS High-Speed SRAM 29

3.1 Overview 29

3.2 CMOS/Bipolar Device Structure 33

3.3 Sense Amplifier Consideration 37

3.4 High Speed 64K SRAM Design 45

3.5 Discussions about Bipolar Application in CMOS Circuits 54

3.5.1 N^+ Buried Layer 54

3.5.2 Gate Delay 55

3.5.3 Sensitivity to a Small Signal 59

3.5.4 Current Driving Capability 60

3.6 Summary 62

4. Design of a 256Kbit CMOS EEPROM 65

4.1 Overview 65

4.2 Analysis of Single Polysilicon EEPROM Cells 69

4.2.1 Erase Operation 69

4.2.2 Program Operation 76

4.2.3 Transient Analysis 78

4.2.4 Experimental Results and Discussions 80

4.3 Cell Structure used for the 256Kbit EEPROM 86

4.4 Open Bit-line Structure 95

4.5 Erase/Program Control Circuits 99

4.6 Performance and Results 106

4.7 Summary 108

5. Design of an Application Specific Memory IC for LSI Function Testing 111

5.1 Overview 111

5.2 System Design 113

5.2.1 Block Diagram 116

5.2.2 Pin Drive 116

5.2.3 Sequencer 119

5.3 On-Chip Memory Design 122

5.4 Circuit Descriptions 129

5.4.1 Short-Circuit Protection 129

5.4.2 Variable Data Acquisition 129

5.4.3 Counter 131

5.5 Design Environment 134

5.6 Results 137

5.7 Summary 142

6. Conclusions 145

6.1 Individual Device Results 147

6.2 High-Speed Capability 149

6.3 High-Functionality 152

6.4 Design Methodology 153

6.5 Final Thoughts 154

References 157

List of Figures 163

Publications 170

Acknowledgements

Without the help of many people, this thesis would not have been possible. Professors Keikichi Tamaru, Shuzo Yajima, Akio Sasaki, and Ryohei Itatani of Kyoto University deserve special mention for reading drafts of this work and providing valuable feedback. This thesis has benefitted from many helpful discussions I have had in Toshiba Semiconductor Device Engineering Laboratory. I would like to thank Drs. Satoshi Shinozaki, Kazuyoshi Shinada, Shinji Saitoh, Susumu Kohyama, Kohichi Kanzaki, Hiroshi Momose, Junichi Tsujimoto, and Naohiro Matsukawa for their useful discussions. I would like to thank Professors Mark Horowitz, and Robert Dutton for their encouragement and discussions during my stay at Stanford University. Finally, I would like to thank Drs. Tetsuya Iizuka, Yoshio Nishi, and Osamu Ozawa, for their continuous support and encouragement through the work.

Chapter 1

Introduction

The origin of solid-state circuits can be traced to the transistor invention by J.Bardeen, W.Brattain, and W.Shockley in 1948. In 1959, R. Noyce and J. Kelby changed the technological and economic profile of the world, when they independently invented the planer integrated circuits (IC). Since 1961 the number of transistors that can be successfully fabricated on a single chip has doubled almost every year. ICs were roughly classified as SSI (Small Scale Integration), MSI (Medium Scale Integration), and LSI (Large Scale Integration), according to a number of transistors on a chip. Nowadays, in order to express the integration beyond LSI, which means the contents of over ten thousands of transistors, people create new words, "VLSI" (Very Large Scale Integration) and "ULSI" (Ultra .).

SSI and MSI in 1960s were mainly manufactured by bipolar transistors, because of the fabrication difficulty of MOSFET (Metal Oxide Insulator Field Effect Transistor). For the stable digital operation, a logical gate progressed in TTL (Transisitor Transistor Logic). TTL input/output specifications were proposed and established. This standardization expanded TTL application to almost all the electric equipment. Many kinds of TTL IC were manufactured to meet the market's needs [1], forming the TTL family.

In 1970s, MOS began to be practically utilized as a

basic IC component as a result of the detailed study of silicon di-oxide (SiO_2). The technology for integrating electric components on a chip has been drastically progressed since that time. Most of the bipolar MSI and LSI were replaced and redesigned by MOS. There were some reasons for it. First, an MOS gate needs less operating current than a bipolar gate, which is necessary for achieving larger scale integration. Second, it is fabricated by less process steps, which might reduce IC cost. Moreover, the scaling law predicts that a smaller MOS would offer higher switching speed with lower power dissipation, approximately proportional to the dimension. It leads the technological enhancement of reducing the minimum feature size of an MOS.

On the other hand, LSI consist of bipolar transistors still survives. In the high speed application, such as main frame computers, ECL (Emitter Couppled Logic) has been widely used. Also, the newly structured device, IIL (Integrated Injection Logic), proposed in 1974, is found suitable for an LSI component, because of its small power dissipation. Some investigations have been performed on this bipolar device [2,3], and it has been applied to the analog and digital mixed IC, for example, TV and VTR control. Besides ECL and IIL, it is to be noted that most of the electric equipment still uses a lot of TTL SSI and MSI in order to "glue" various kinds of MOS LSI. For this reason, almost all the MOS LSI are taking TTL compatible input/output specifications even now.

Before 1980, most of LSI were designed by Nchannel MOSFET(NMOS) as it could be made by a simple fabrication process and could have higher switching speed than Pchannel MOSFET(PMOS). However, this tendency has been changing, recently. If one wants to integrate millions of transistors on a chip, he would find that even little current for one transistor causes to break the package power limit. In addition, larger design margin for each circuit would be required to guarantee the whole function. Saving the design labor becomes a key issue, as the numbers of the total components increase, and the characteristics of each component made by more sophisticated device miniaturization process fluctuates. From the point of view, VLSI designed by CMOS (Complementary MOS, composed of NMOS and PMOS) has been gaining the populality for VLSI and ULSI. CMOS is requied no power in steady state, and is less affected by the transistor dimension, although it is still slower and naturally needs more process steps than only NMOS. Therefore, many circuit and device technologies to overcome the problems have been proposed and applied to the CMOS LSI.

One of the most important LSIs which lead the technological development of both fabrication process and circuits is a memory. The demand for higher performance memories, in other words, memories having larger bit density, faster access time, lower power dissipation and higher reliability has still been increasing widely. Memories are classified as RAM (Random Access Memory), and

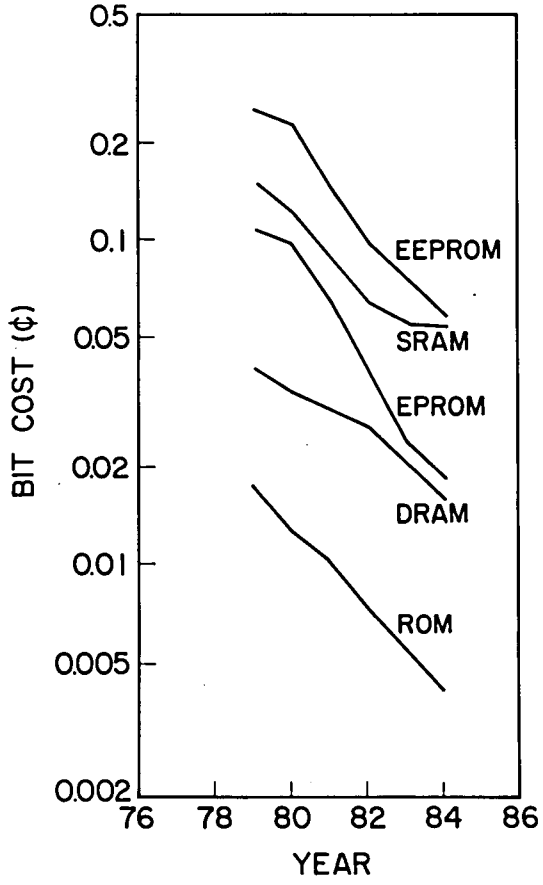


Fig. 1.1 Memory bit cost.

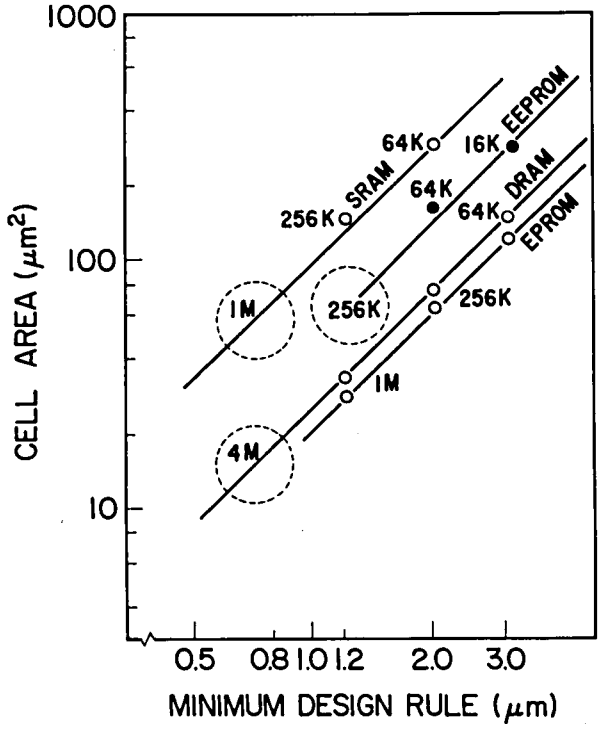


Fig. 1.2 Memory cell area as a function of minimum design rule.

ROM (Read Only Memory) by the data non-volatility. RAM is divided into SRAM (Static RAM), operating asynchronously without reflesh, and DRAM (Dynamic RAM) operating synchronously with periodical reflesh. ROM is divided into Mask ROM, UV-EPROM (Ultra Violet Erasable Programmable ROM) and EEPROM (Electrically Erasable PROM). They offer different bit cost, and turn-around time (TAT) which is defined as the time from programming data to reading data. Bit-cost is decreasing year by year in each memory, as shown in Fig. 1.1. It is mainly caused by the cell size reduction, which is enabled by technological improvement. Figure 1.2 shows the cell size as a function of the design rule. Dashed circles indicate the currently developing devices.

Among various kinds of RAMs, bipolar RAM (not shown in either Fig. 1.1 or 1.2) has been very high-speed, while it has relatively smaller bit density, and consumes more current. The market is relatively small and is limited to a special use such as the cache in the main frame computer. Static MOS RAM has aimed low power dissipation and ease-of-use with medium speed and medium density for portable electrical equipment. Dynamic RAM has achieved the highest bit density and the lowest bit-cost, which has been used for main memory of computers. However, nowadays, the application of RAM reveals a lot of diversities and the borders become ambiguous. In other words, there is a great demand for CMOS SRAM with high speed as well as the bipolar and with high density and low cost as MOS DRAM, keeping its

low power attribution. The application is spreading into the cache around the micro processors. Similarly, many DRAM application methods have been proposed to make it fast and ease of use virtually, resulting in the invasion to previous SRAM market. Consequently, especially for SRAM, the circuit and device innovation is required to satisfy the market's needs, taking advantage of its properties.

On the other hand, among the non-volatile memories, EEPROM has seemed to be an ultimate memory, as it can be read and written electrically on the board. Every market research has indicated that the amount of the world-wide production would become the largest in some day. But, it has not. Actually, the bit-cost is the highest among memories and the development of higher density EEPROM is the slowest as indicated in Figs. 1.1, and 1.2. Of course, it owes to technological difficulties such as a thin oxide formation, and high voltage transistor structure. But, an idea may break through the hedge.

Besides the trend of standard memories having rather simple architecture, there will be proposed several ASMIC (Application Specific Memory IC) to meet the diversities of memory application. The established fabrication process verified by the standard memories will be used to realize ASMIC. The peripheral or interface logic will be combined into the memory to eliminate several on-board ICs and to improve system performance, by avoiding interface delay. So, circuit or architectural technology to combine function units with an in-system memory enough to meet individual

specifications is required.

Revealing such diversities as mentioned above, however, each memory always takes the steady step toward high-speed, high-density and high functionality. Recently, processors have been able to work in higher clock rate, and to support more memory address field. Now, the performance of computer system is said to be limited by the memory band width. One way to solve the memory problem is, of course, to achieve high-speed by itself. The another way is to promote the memory value, in other words, to be more intelligent, to be more functional, or to have larger bits in it. Generally, in the board level, IC is demanded to drive long capacitive and inductive wires in addition to the capacitive load of receiver ICs. Taking account of noise and voltage fluctuation, the system design needs to take large operating margin, which sacrifices the operating speed. On the other hand, from the IC designer's point of view, internal signals having rather small drivability should be amplified in the last stage to drive the external load. Or, to receive noisy input signal accurately, the input buffer should be designed small input impedance with low pass filter. Sometimes, a certain amplification and level shift, fitting to the internal signal level are required. As a result, the percentage of the input and output buffer delay in a total access time becomes quite high. So, integrating more functions into a chip becomes more important. Sometimes, the intelligence in a memory saves the load of processor tasks.

From the above circumstances around memories, this thesis will describe four memory design works, an ECL RAM, a Bi-CMOS RAM, a CMOS EEPROM, and an ASMIC for LSI functional testing, focussing on CMOS technology. All aim at improving the access time, increasing the bit density, and integrating more functions on a chip.

In order to achieve high-speed and high-functionality, these four devices took different approaches each other. The ECL-RAM used high performance components such as the oxide isolated bipolar transistor, Schottky barrier diode, double level metal, and polysilicon resistor. It allows much current for charging or discharging word- and bit-lines. As a result, it achievs 4.6ns access time with 340mW power dissipation. The design of the 256bit ECL-RAM [6] will be presented in Chapter 2. The design approach for the high speed memory was succeeded by the Bi-CMOS (Bipolar and CMOS) RAM in the following Chapter.

High speed CMOS SRAM needs to realize quite low standby power, to take TTL compatibility, and to integrate more bits in a chip. The newly developed Bi-CMOS structure enables co-existence of NPN transistor and CMOS without any additional process steps. By using NPN transistors for bit line sense amplifiers, the 64K Bi-CMOS RAM works in 28ns address access time, with the power dissipation of 225mW active and 100nW standby. The access had never been achieved by only the CMOS technology. The detail of the design [7-10] will be described in Chapter 3.

EEPROM has an advantage of non-volatility, different

from RAM, but had been suffered from the slow access and programming time, which came from its device structure. By controling the circuits' operation by internally generated clocks and taking distributed sense amplifier architecture, it enables 150ns access time. For the fast programming, the page-mode scheme was applied and enabled 225us programming time per byte apparantly. A unique single polysilicon cell enables to integrate the largest EEPROM bits on a chip. The design of the high performance 256Kbit EEPROM [11-13] will be covered in Chapter 4.

The high speed and high functional memory design is required for an ASMIC. One chip functional tester, realized by a novel architecture belongs to ASMICs. It generates test vectors by the on-chip memory data, receives DUT (Device Under Test) output data, saves them into the memory, and compares them with the expected data. The performance of the on-chip SRAM determines its maximum operating frequency. To increase the test vector rate, it adopted pipelining technique. 4 words are read at one time from the memory, but the output data comes out one by one. It virtually gains over 10MHz vector rate by 3um design ruled CMOS, and 16MHz by 2um. The design of the one chip tester [14,15] will be shown in Chapter 5.

Chapter 6 summarizes the contribution of this thesis. Areas for further investigations are also described.

10 項欠

Chapter 2

Design of a 256bit ECL RAM Using High-Speed Bipolar Technology

2.1 Overview

The high speed computer system is usually organized by fast processor, fast cache, medium speed memory, and slow mass storage. So, considerable effort has been applied to the development of fast and complex ECL-memories [21,22], motivated by the continuing need to bridge the speed mismatch of today's fast logic circuits and low-cost high capacity MOS RAMs. Especially, the development of ECL gate array technology makes it easy to design application specific ECL processors, stimulating the production for the high speed ECL memory market.

As it has been used for the bridge device, such as cache, local memory, and a register file for a high performance computer peripheral, the improvement of ECL RAM has been made focussing rather on speed than on bit density. The shallow junction bipolar transistor with an oxide isolation having high cut-off frequency, has been applied to the memory. And then, Schottky barrier diode (SBD) is introduced in order to avoid transistor saturation even in the large current driving region. Now, it becomes important to form a polysilicon resistor to eliminate the parasitic capacitance and to save cell area. However, the controlability of the polysilicon was not sufficiently

good, because the resistance value is sensitive to the doping concentration and the grain size. In addition, for the ohmic contact between the lightly doped polysilicon and the metal, N^+ diffusion should be carried out. But the heavily doped region diffuses laterally, which results in the drastic decrease of the resistance value. As the N^+ diffusion process step is common to the N^+ formation for the emitter, it affects not only the resistor but the base width which directly determines transistor gain and cut-off frequency. So, the process condition should be optimized to satisfy both of the resistance and the base width.

This Chapter presents the development of a 256 bit ECL RAM, integrating high-speed bipolar technology such as oxide isolated bipolar transistors, Schottky barrier diode, polysilicon resistors, the double level metal, and high-speed circuit configurations. Section 2.2 describes the cell design and fabrication process steps. In Section 2.3, the circuit design and the results will be shown. Finally, the work will be summarized in Section 2.4. The circuit design conception of high-speed bipolar memory will be discussed as well, and it will be applied to the design of a Bi-CMOS RAM in the following Chapter.

2.2 Cell Design and Process Sequence

The circuit design of ECL RAM started at memory cell design, that is to say, how to reduce the cell area and cell holding current, I_H. And both requirements can be achieved by using polysilicon resistor as a collector load. In comparison with the diffusion resistor made in the active region, the polysilicon resistor can be formed on the silicon di-oxide above the isolation region, which results in saving cell area. The sheet resistivity (ρ_s) of the diffusion region, applicable for a large register, is limited by depletion region effect and thermal co-efficient. Much higher diffusion resistor needs quite longer length. Polysilicon resistors can achieve high resistance value without sacrificing the dimension. So, I_H is saved by increasing the resistance, keeping the cell voltage difference between the two collector nodes (holding voltage) constant. Furthermore, in comparison with the diffusion resistor, the polysilicon resistor has a less parasitic capacitance, which is necessary to achieve high speed. The cell layout, taken for the 256bit ECL RAM is made by a multi-emitter flip-flop, as shown in Fig. 2.1. The pattern layout, the equivalent circuit, and the cross section view of a cell are illustrated, also. The NPN transistors are isolated each other by silicon di-oxide region, which reduces the parasitic collector - substrate capacitance. The collector node was clamped to the base node by Schottky barrier diode for keeping every NPN transistor to operate in the non-saturated region. The cell

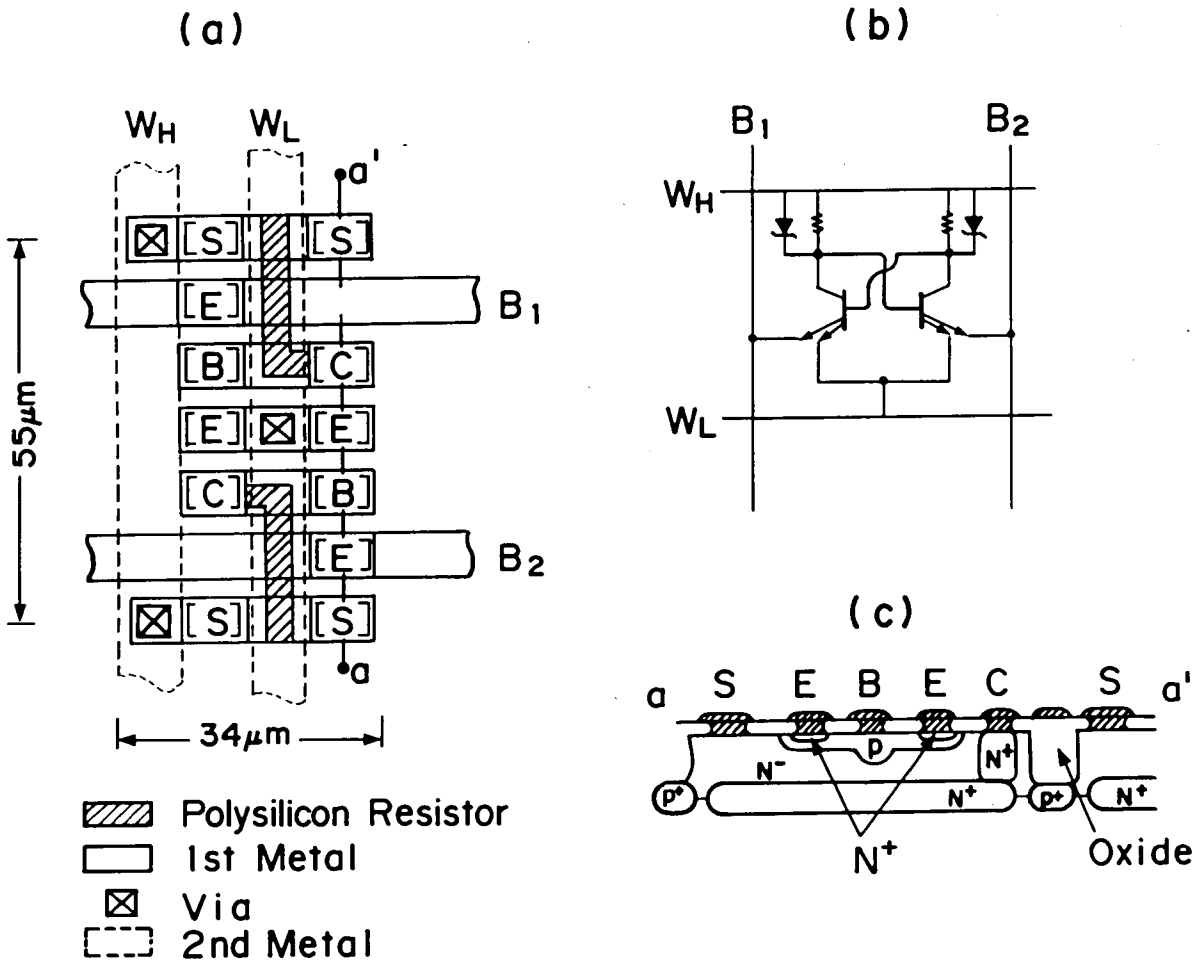


Fig. 2.1 (a) Cell layout, (b) Equivalent circuit, and (c) Cross section view of ECL memory.

area was 55*34um by using 3um minimum design rule.

The point of making this type of cell was the controlability of polysilicon resistor load. The process steps to make it was as follows;

(1) Undoped polysilicon deposition (200nm thick) by 600°C LP CVD on the thermal oxidized silicon substrate.

(2) pattern etching

(3) Light As^+ (Arsenic) implantation for the intrinsic resistor. The dose was 10^{14}-$10^{15}cm^{-2}$ at 100keV, depending on the resistance value, taking controlability into account.

(4) Heavy As^+ implantation to the resistor contact region, forming the emitter, simultaneously.
The dose was $10^{15}cm^{-2}$ at 60keV.

(5) 1000° annealing.

The sheet resistivity, ρ_s, as a function of the anealing time is shown in Fig. 2.2. ρ_s is almost independent of the anealing time, if the dose is ranging from 10^{14} to $10^{15}cm^{-2}$. So, ρ_s dependence on the As^+ dose can be drawn as Fig. 2.3. This figure shows that ρ_s is quite sensitive to the fluctuation of the As^+ dose.

For the lightly doped polysilicon, ranging from 10^{14} to $10^{15}cm^{-2}$ dose, it is too difficult to make an ohmic contact to the Aluminum. The resistor needs another heavy dose ion implantation (As^+) on the contact area. The sheet resistivity of As^+ ($10^{15}cm^{-2}$ dose) is about $50\Omega/\square$, negligiblly smaller than the ρ_s of lightly doped

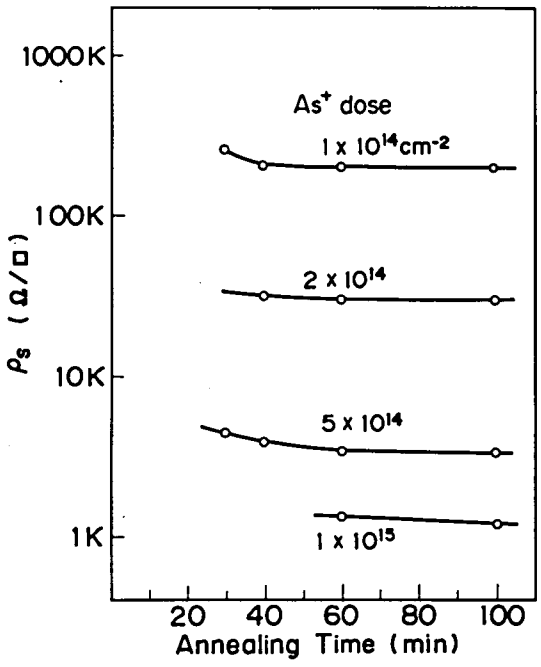


Fig. 2.2
Annealing time dependence of sheet resistivity.

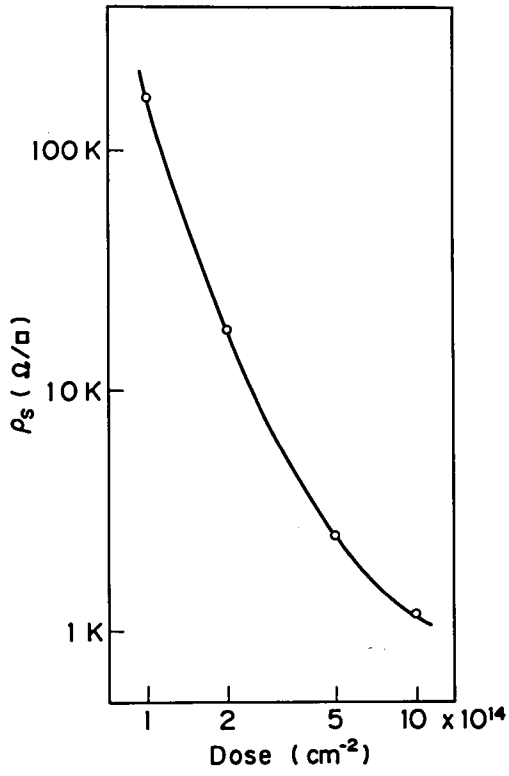


Fig. 2.3
As+ dose dependence on sheet resistivity.

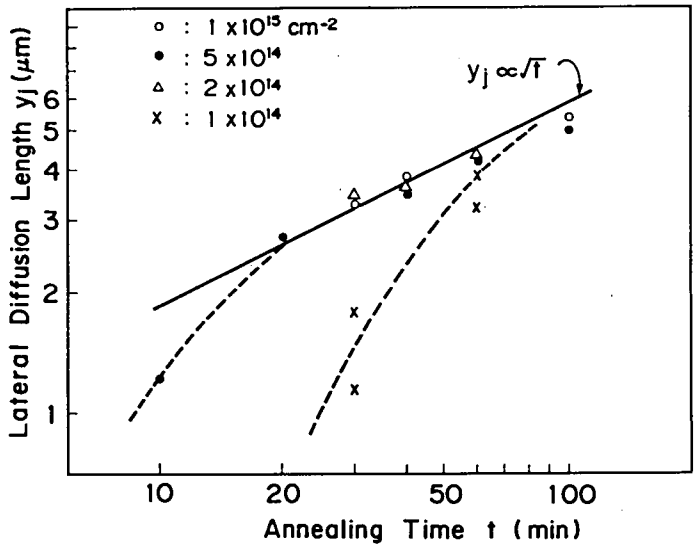


Fig. 2.4 Annealing time dependence on lateral diffusion, y_j

polysilicon. So, the value of the polysilicon resistance, R, is approximately given by the following equation,

$$R = \rho_s * (L - 2y_j) / W \qquad (2.1)$$

where L and W are the length and width of the resistance. y_j is the As^+ lateral diffusion length from the contact area. Although ρ_s is almost independent of the anealing time, y_j is dominated by it. By measuring the several lengthes of resistances, y_j can be calculated statistically. Figure 2.4 shows the anealing time dependence of the lateral diffusion length, y_j for various kinds of As^+ dose. The lateral diffusion length is approximately proportional to the square root of the anealing time, if the anealing time exceeds 60min. This result indicates that the diffusion length of As^+ almost follows the diffusion equation, although the co-efficient is much larger than that of silicon substrate. And the value can be controlled by the anealing time, if the anealing time is long enough.

The anealing condition determines another important parameter, the base width of NPN transistor, which affects the characteristics of NPN transistor. As the emitter and the contact of polysilicon resistance are made simultaneously, the anealing time should be determined carefully.

The resistance variation on a wafer was examined. Using the polysilicon resistor of L/W=22um/3um with $7.5*10^{14}cm^{-2}$

dose, the average and deviation were 9.84kΩ and 1.05kΩ, respectively. On the other hand, those of the diffusion layer resistor of L/W=20um/3um were 7.14kΩ and 1.02kΩ. So, the polysilicon resistance, having the almost same deviation as the conventional one, was successfully made.

This technology was applied to the fabrication of the 256bit ECL RAM. The whole process steps are summarized as follows;

(1) The epitaxial layer growth on P^- type silicon substrate having N^+ buried layer. The thickness was 2.0 um, and the resistivity was 0.2Ωcm.

(2) The dielectric isolation was carried out.

(3) Boron implantation of $10^{14}cm^{-2}$ to form the base region with the diffusion depth of 0.5um.

(4) Arsenic Implantation of $10^{15}cm^{-2}$ to form the emitter region with the diffusion depth of 0.3um.

(5) As^+ implantaion of $7.5*10^{14}cm^{-2}$ to form the polysilicon resistor. The L/W was chosen as 22um/3um.

(6) 75min anealing at 1000°C. The β (the gain) was 50.

(7) Evaporation of PtSi on N^- layer for the Schottky barrier diode.

(8) Double layer metalization both by AlSi. Plasma deposited SiN was used for the insulator between the two metals.

2.3 Circuit Design

The circuit block diagram of a 256bit ECL RAM is shown in Fig. 2.5. The circuit design of a memory started at the architectural determination around the cell. The detailed circuitry around the cell is shown in Fig. 2.6. Time dependent internal signal waveforms and bias voltage of the RAM, obtained by the circuit simulation, SPICE, are depicted in Fig. 2.7.

The memory read and write function, adopted for the ECL-RAM is explained as follows;

The row address signal with ECL level (H: -0.8V, L: -1.6V) is received by the input buffer at the time of 0ns. After one out of 16 word-lines (WL) is selected and is elevated from -2.4V to -1.6V, the two base nodes (Cell "H", and Cell "L") of each cell connected to this WL follow it, as shown in Fig. 2.7. By the column address inputs one of the bit- line pairs is selected, for example, CS0 goes to higher than the others. As a result, the read current, I_R, flows only in the selected bit-line pair. As long as those base nodes are lower than the reference level, V_R, (within 1.2ns in Fig. 2.7) both of the I_R flow in the reference transistors (Q_R). But, as soon as the higher base node of the selected cell exceeds V_R, I_R is provided by the ON transistor with higher base voltage through its load resistor. Consequently, current difference between two reference transistors is detected by the sense amplifier. This signal is amplified by the output ECL buffer enough to drive the off-chip load. In the write mode, one of the base

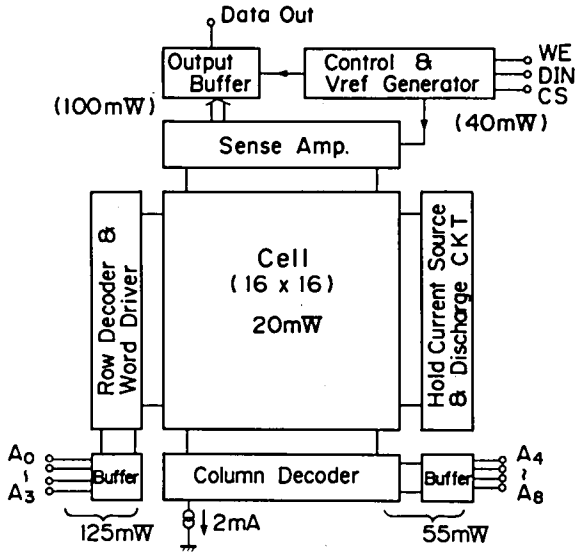


Fig. 2.5 Memory block diagram of 256bit ECL RAM.

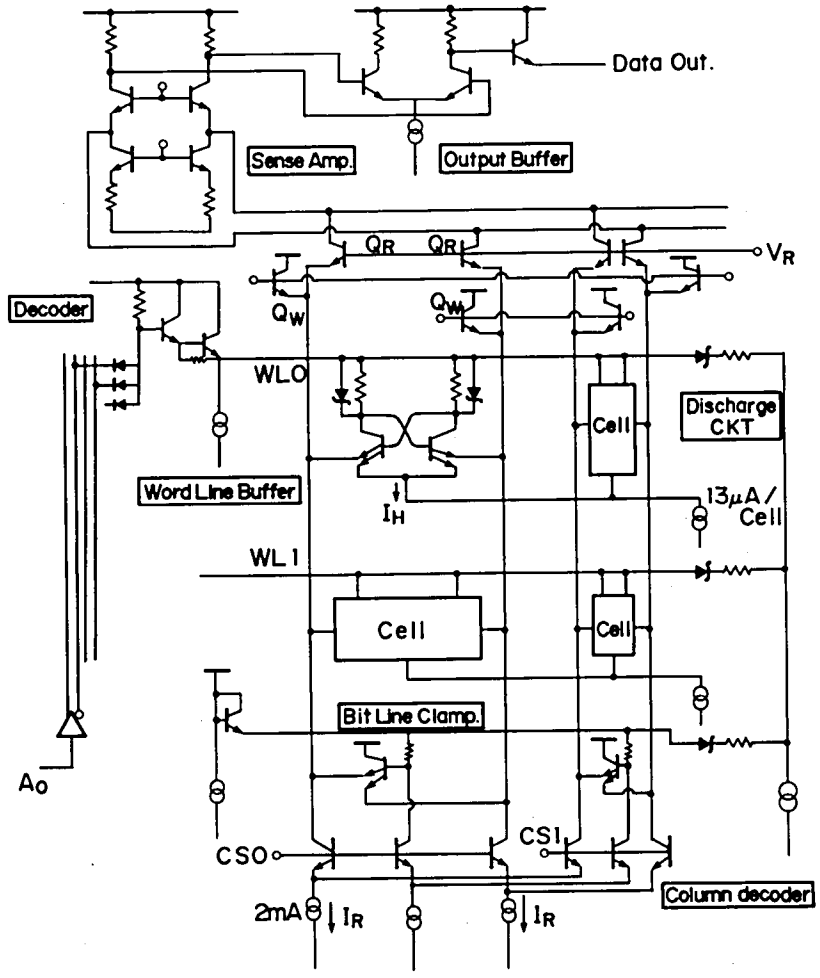


Fig. 2.6 Memory cell peripheral circuitry.

of Q_W is set to "H-write", (-1.6V) and another Q_W is to "L-write", (-2.4V) in Fig. 2.7. By the same selecting sequence of the word-line and bit-line pair as reading, I_R flows only the cell transistor emitter-coupled with the "L-write" base transistor, Q_W, which makes its collector node lower than the other, whatever the transistor may be previously ON or OFF. Unselected cells connected to the same bit-line pair are not influenced, because each cell node is lower than the "L-write" voltage. Unselected cells attached to the same word-line are not affected because no I_R flows.

The design of the read current, I_R, is a key issue to accelerate address access time, because the bit-line delay is almost proportionl to C_{BL}/I_R, where C_{BL} is a bit-line capacitance. Larger I_R gives the other benefit for total delay reduction. The bias current of the sense amplifier, I_S, is chosen equivalent to I_R. Larger I_S results in faster sensing speed, and larger drivability for the next gate. Some extra cascaded gates for amplifying enough to drive the external output load might be eliminated by the drivability.

For the wide operating margin, the holding voltage above 100mV is commonly used for the ECL memory design. I_H should be designed lower to save the total power, while I_R should be larger for the fast access time. However, The value of I_R could take only six times larger than I_H, if one tries to keep the voltage drop of the load resistance within the base -emitter voltage drop, V_f, (0.7V). The

operation in the saturation region degrades the switching speed. In order to break such limitation, a parallel connected Schottky barrier diode (SBD), as shown in Fig. 2.6, was implemented, which clamps the collector voltage to $V_B - V_S$ even in a larger collector current, where V_B is the base voltage and V_S is the SBD cut-in voltage. Note that

$$V_f - V_S >> VCE(sat), \qquad (2.2)$$

hence VCE(sat) is the saturation voltage of the NPN transistors. In this configuration, the row address access time dependence on I_R was obtained by the circuit simulator, SPICE. The access time as a function of $1/I_R$ is shown in Fig. 2.8. It clearly indicates that the bit-line delay decreases as I_R increases. The limiting factor of maximum I_R value, except for the total power consumption, is series parasitic resistance of SBD, R_S. To avoid the saturation, the following condition must be satisfied,

$$I_R < (V_f - V_S)/R_S. \qquad (2.3)$$

From the measurement, V_S was 0.4V, and R_S was 150Ω. So, 2mA was chosen for an I_R. At the design point, the simulation predicted 4.5ns access time, as shown in Figs. 2.7 and 2.8. And from the experimental results mentioned in previous subsection, around 10 kΩ poly resistance could be made in stable. So, the holding current was selected as 13uA.

In the memory, bit-line swing, ΔV_{BL}, was designed as;

$$\Delta V_{BL} = V_Bh - V_R = 0.3V, \qquad (2.4)$$

where, V_Bh is the higher level of the cell base node. Note that ΔV_{BL} is about half of the base voltage swing of the selected cell. There is a voltage drop from the word-line

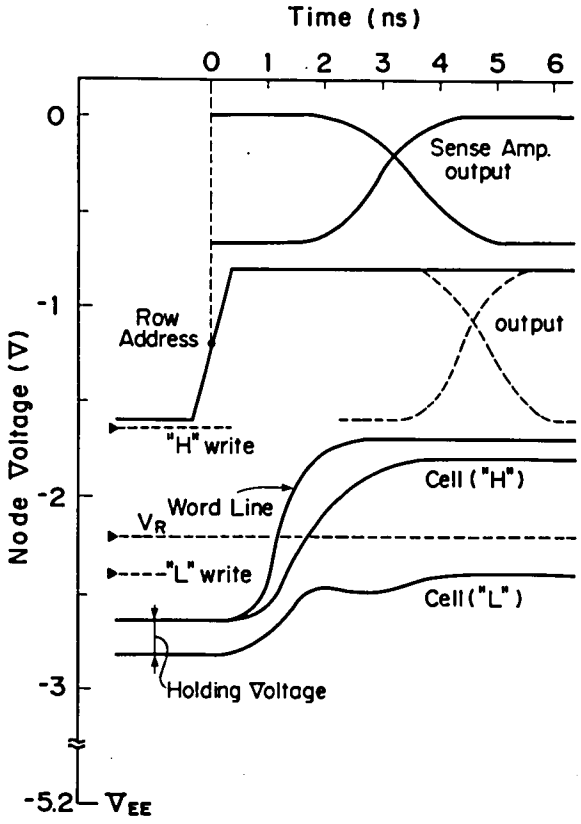


Fig. 2.7 Time dependent signal and bias voltage of internal nodes.

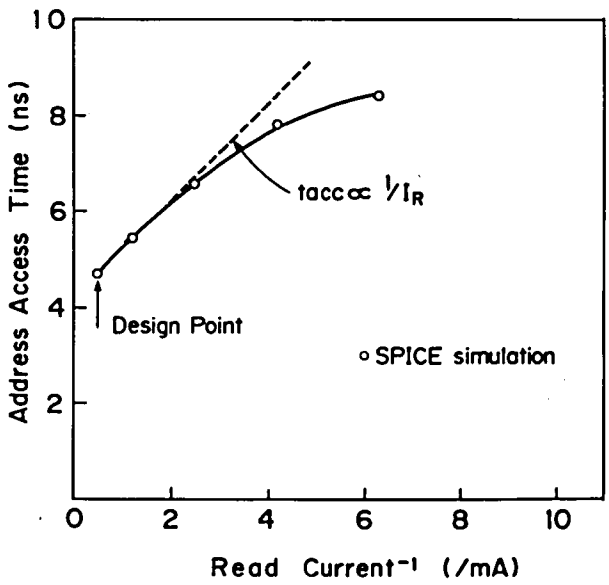


Fig. 2.8 Read current dependence of access time.

to the higher base node caused by I_R/β current flow.

Besides the bit-line delay reduction, quick rise and fall of the word-line is another important issue. As the word-line was made by the second level metal, the RC delay is not so significant, as long as the drivability is large enough. So, the word-line driver of the Darlington Configuration was incorporated to raise it fast. However, Darlington Configuration could not accelerate the word-line pull-down. As the multi word-line selection at the moment of the address transition may degrade the address access time, the discharge circuit shown in Fig. 2.6 was implemented. The discharge current was pulled from the highest level of the word-line, making it lower quickly.

A microphotograph of a 256 bit ECL RAM is shown in Fig. 2.9. The minimum emitter size of the NPN transistor is 3um*6um. The first metal line and space are 5um and 3um, respectively. Those of the second metal are 7um and 5um. Exploiting 3um minimum feature size, the die size is 2.0mm*2.3mm. As shown in Fig. 2.10, the row address access time was obsereved as 4.6ns. The column access was 4.4ns under the 340mW power consumption. The measured access time agrees with the simulation. The characteristics of this RAM is summarized in Tab. 2.1.

The 256 bit ECL RAM was successfully obtained, by using the high-speed bipolar technology.

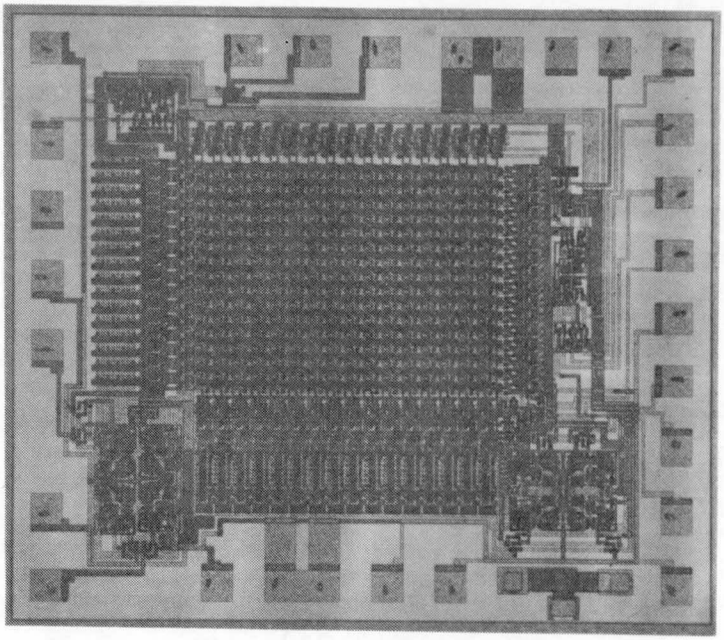


Fig. 2.9 Microphotograph of 256bit ECL RAM.

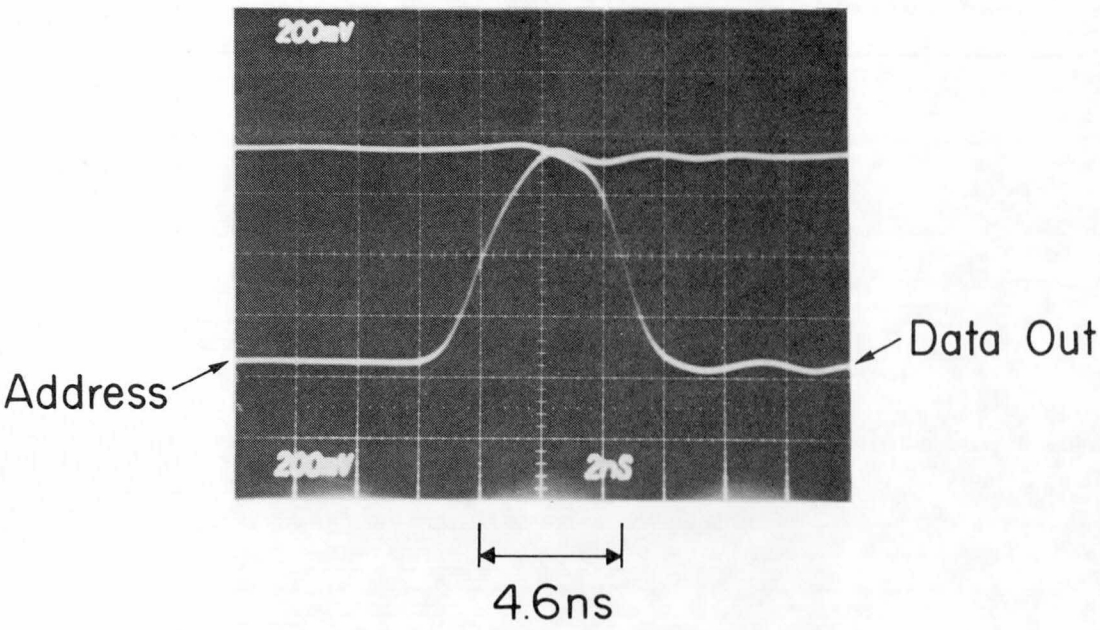


Fig. 2.10 Waveforms of address input and data output.

Tab. 2.1 Characteristics of 256bit ECL RAM.

Organization	256*1bit
Operation	Fully asynchronous
Technology	Double Level Metal Bipolar with SBD and poly resistors
Cell size	55*34um²
Die size	2.0*2.5mm²
I/O interface	ECL
Address Access Time	4.6ns(typical)
Active Power	340mW(V_{cc}=5V)
Holding Current	13uA
Read Current	2mA

2.4 Summary

Using the oxide isolated bipolar transistor, double level metal, Schottky barrier diode, polysilicon resistors, and high-speed circuit configurations, a 256bit ECL RAM has been developed. It offers 4,6ns row, and 4.4ns column address access time, under 340mW power consumption. The read current was 2mA, which was determined by the bit-line discharging time, and the parasitic series resistance of Schottky barrier diode. And the hold current was set to 13uA, designed from the polysilicon resistance value.

The polysilicon resistor is a key technology to realize such a high speed memory. Lightly doped polysilicon enables higher value of resistance, avoiding junction capacitance in addition to the cell area saving. Both of the ion implantation dose and anealing time affect the sheet resistivity. Moreover, heavily doped region is required for the ohmic contact to aluminum. The lateral diffusion of this high concentration region makes the resistance value small.

It is found that the lateral impurity diffusion on a polysilicon almost follows the same equation as the one on a bulk, as long as the anealing time exceeds a certain value which is dominated by the ion implantation dose. But, the diffusion co-efficient is quite large. In case of the 256bit RAM, the parameters of $7.5*10^{14}cm^{-2}$ dose for the intrinsic resistance, and $1*10^{15}cm^{-2}$ dose for the contact region with the metal, and 75min. annealing time were carried out. The average value of L/W=22um/3um resistance

was 9,84kΩ. The chip-to-chip standard deviation was 1.05kΩ which was comparable to the resistance made by the diffusion on a bulk. Under the same diffusion condition, the emitter depth was 0.3um, resulting in the 0.2um base width. The NPN transistor offers 50 current gain, and 3.8GHz cut-off frequency.

As a result, 4.4ns worst-case access time, the fastest in the world, was obtained by this ECL RAM. It is proved that polysilicon resistors are controlable and useful for the high-speed bipolar RAM.

Besides peculiar process steps for bipolar LSIs, there are some circuit design conceptions to be noted, which are listed as follows;

1. Using small amplitude signals as 0.5 - 0.7V cuts off the charging and discharging delay time of long capacitive lines.
2. The stable cut-in voltage, V_f, independent of the process fluctuation, enables high-speed signal detection.
3. In the bipolar transistor, the current increases exponentially with voltage, which enables to design high gain amplifiers.
4. High drivability buffers such as the emitter follower and the Darlington Configurations can be used.

Some of them can be applied to CMOS circuits. The detailed study will be shown in the following Chapter.

Chapter 3

Design of a 64K Bipolar-CMOS High-Speed SRAM

3.1 Overview

CMOS circuits have inherent advantages of low power and large noise margin. Several high density and low power random access memories (RAM) with access times in the range of 65 to 80 ns have been intensively developed [16-18]. They have been used for the medium speed MOS micro-processors. Recently, as peripheral memories of high-speed bipolar (TTL) processors, high-speed TTL compatible CMOS memories, comparable to the NMOS and ECL memories have been demanded, keeping CMOS properties of low power and high density. The TTL compatible NMOS and bipolar ECL memories have already achieved 30ns or less [19-21], although they have exhausted as much power as about 1W with less than 16kbits.

First, in order to achieve fast access time, the delay time reduction in long capacitive lines such as word lines and bit lines is of vital importance, all over the memories. From the circuit innovation point of view, the use of the internal clock has been considered to be one of the most promising methods [19]. The clock is generated at any transitions of address changings. It precharges and equalizes the bit line pair, or activates the sense amplifiers. However, as there are a lot of lines with parasitic capacitance to be charged, the clock should be

amplified enough by the series connections of several buffers, which becomes the cause of delay. In addition, too long precharge period leads the access time loss. So, the clock pulse width should be optimized. However, generating stable width clock under any input condition needs stable device characteristics. Consequently, this approach requires both adjustment of complicated clock timings and precise control of device fabrication.

CMOS process also has many compatible steps with bipolar device process. several approaches to integrate both CMOS and bipolar devices on a same chip have been reported. The suitability of bipolar devices for analog parts in analog-digital VLSI systems has been discussed. Nevertheless, the application of bipolar devices in CMOS memories has been mostly limited to output buffers [23], or level shifters [24], where the collectors are biased to a fixed level. As an output buffer, an NPN load and NMOS driver type buffer has been commonly utilized. Generally, output buffers on CMOS LSI, which often drive CMOS devices, are required full swing operation. If the output voltage settles in the middle between the GND and the supply voltage, V_{DD}, both PMOS and NMOS of the next CMOS input gates are driven to ON, allowing the idle current flow. The amount of current consumption in a whole system increases, as the number of connected devices increase. It is difficult for NPN load to elevate the signal to V_{DD} without additional elements. In addition, if the interfaced device pulls much emitter current from the buffer, the

collector or the well voltage drops. It may induce latch-up or breakdown of the whole system. Therefore, it might be preferable to make the output buffer by the CMOS. Of course, the situation would be changed, if the formations of the buried layer with the epitaxial growth and the deep N^+ between the buried layer and the collector contact to reduce the internal collector resistance are acceptable. However, even though CMOS-bipolar technology is applied, process simplicity will be still required for low-cost, as well as for no degradation of each device.

This chapter describes the design of a TTL-compatible 64K CMOS SRAM with a new Bipolar-CMOS technology, adding some discussions about the combined technology. To improve the bit line delay, the bit line swing was reduced to 0.7V. The value was common to the conventional ECL memories, but has not applied to pure CMOS memories. Because an MOSFET generally needs large gate voltage swing to obtain higher drivability. A high sensitive differential amplifier composed of emitter coupled NPN transistors was newly adopted to enable the small signal detection, and to tolerate the characteristic variation of a large number of cells, while CMOS type memory cells were taken because of the small cell-size and low power stand-by. Besides sense amplifiers, the NPN transistors are applied to the bit-line precharge circuits, and the voltage regulators, taking advantage of their large drivability and the stable cut-in voltage. However, the application to the output buffer was given up, because of the large collector resistance and the

full swing requirement as an output voltage. Throughout the design of the Bi-CMOS memory, the design conception of the high-speed bipolar memory which has been mentioned in Chapter 2, was applied.

The following section describes the new CMOS-bipolar device structure and fabrication process. The characteristics of the NPN transistor made by the process will be given. In Section 3.3, a method for the bit line delay reduction will be proposed. Two types of amplifiers, one made by the bipolar and the other by CMOS, will be compared. The experimental results will be introduced. Section 3.4 is a precise description of a high-speed static RAM design. The application of the bipolar devices in CMOS LSI will be discussed in Section 3.5, including Bi-CMOS technology with the N^+ buried layer, although the structure has not been applied to the RAM. The work will be summarized in Section 3.6.

3.2 CMOS/Bipolar Device Structure

An N-well CMOS process inherently realizes collector isolated vertical NPN transistors without an epitaxial growth step, as shown in Fig. 3.1. With this structure, various bipolar circuit techniques, using NPN transistors most of whose collectors are not fixed to a constant voltage, are applied to CMOS LSI design.

For high performance and most cost-competitive memories, a fabrication process was demanded to achieve the following goals; no additional process step should be used. The values of the current gain, β, and the cut off frequency, f_T should be comparable to those made by the conventional bipolar process. The internal collector resistance, r_c, should be reduced as much as possible.

The N^+ source/drain of NMOS and P^+ of PMOS formations realize the emitter and the external base region of NPN transistors, respectively. This double base structure (separating internal P^- and external P^+ base region) is important to reduce the base resistance. As the impurity profile for the internal base region dominates the performance of NPN such as β, f_T, and VCEO (the emitter-collector break down voltage), it should be carefully formed. Lower impurity concentration results in higher β, f_T, but lower VCEO and larger base resistance. On the other hand, as the gate length of NMOS is scaled down, a higher impurity concentration below the channel region is needed to avoid the source-drain punch-through break down. Therefore, the same boron implantation process to the NMOS

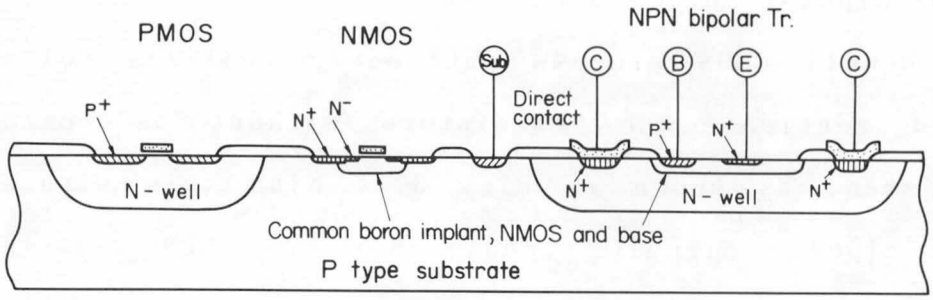


Fig. 3.1 Schematic cross section of the N-well CMOS-bipolar transistors.

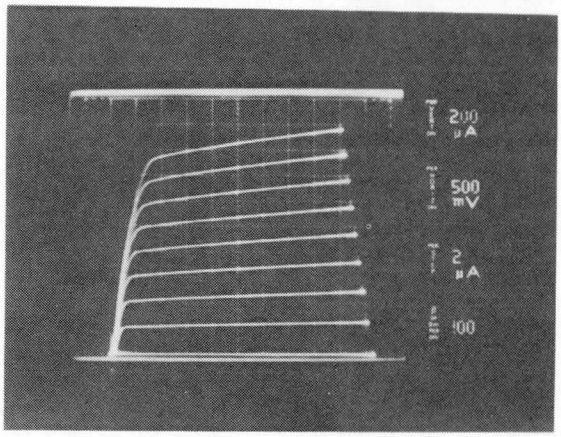


Fig. 3.2 DC characteristics of the NPN transistor.

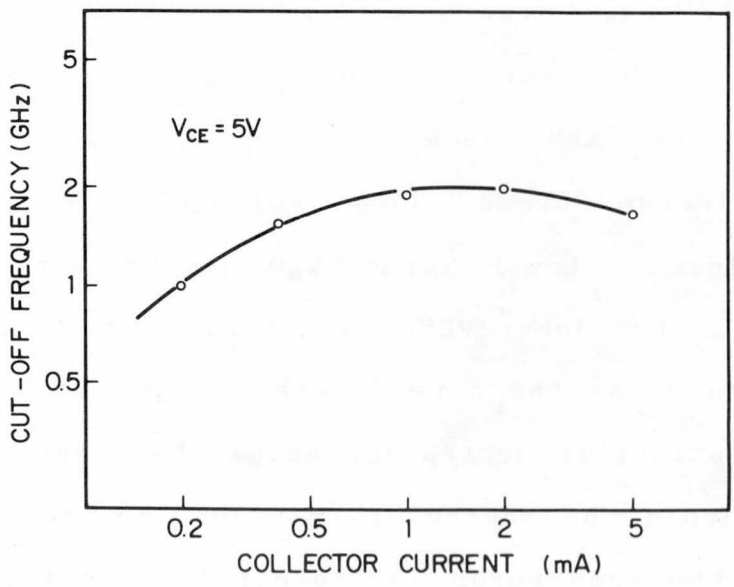


Fig. 3.3 Cut-off frequency of the NPN transistor as a function of the collector current.

channel can be applicable to the formation of the NPN internal base. Actually, in the case of a 1.2um NMOS gate length, the optimized implant dose exists at a range of $2-3*10^{12}cm^{-2}$. In this case, VCEO was about 8V, which guarantees safe application to the internal circuit, because the maximum voltage applied to the memory is specified as 7V (5.5V in the operation). And with this dose, β was ranging from 80 to 200. Even the conventional bipolar process ensures 50-200 as β. Therefore, the NMOS channel and the base regions are formed simultaneously, as shown in Fig. 3.1. The deep N^+ diffusion for the collector region, which is quite important to reduce the horizontal resistance is made by the same process step as the formation of buried contact in standard MOS processes. The N^+ impurity is diffused from the phosphorous doped polysilicon, which enables deeper N^+ diffusion than NMOS source. This collector surrounds the emitter region to reduce the resistance more. Consequently, the structure shown in Fig. 3.1 can be fabricated without any additional process steps than the standard CMOS process. It can give no degradation to either CMOS or NPN transistors.

The photograph of DC characteristics of NPN transistor is shown in Fig. 3.2. The r_c value is reduced to about 600Ω. The base width, the emitter size, and the occupied area of this bipolar transistor are 0.35um, 4.5*7.5um, and 30*37um, respectively. As we stressed on no additional process from the standard CMOS, the transistor does not have a polysilicon emitter structure, which is standard in

the current bipolar process. As a result, the emitter width of 4.5um was determined by the minimum contact width plus the tolerance for the mask alignment between the contact and N^+ emitter. The structure needs larger transistor area, although it easily takes an emitter contact than the other. The cut-off frequency, f_T, as a function of collector current, I_C, is shown in Fig. 3.3. The maximum f_T of 2GHz was obtained at I_C=1mA, which is comparable to the one made by the conventional bipolar process with PN junction isolation. The characteristics make it possible to apply the NPN transistors to a high-speed static memory, which will be described after the next section. A lightly doped drain NMOS structure (LDD) has also been developed in order to reduce the small geometrical effects such as hot electron injection, and the break down voltage lowering. This structure is also shown in Fig. 3.1. With the technology, the break down voltage of a 1.2um gate-length NMOS was improved above 8V.

3.3 Sense Amplifier Considerations

Typical schematic diagram of CMOS SRAM using 6 transistor memory cell is shown in Fig. 3.4. The word/bit-line selection and the read/write scheme of CMOS memory is quite different from that of the ECL memory, shown in Fig. 2.6. In order to obtain larger drivability, the NPN transistor needs larger base current, while the MOS needs larger gate voltage. In the ECL memory scheme, several voltage levels, classified by the unit of diode forward bias (0.8V) were exploited. But, in the CMOS, all the levels besides bit-lines are set to GND, or V_{DD}. For example, the word-line amplitude was 0.8V for ECL, as shown in Fig. 2.7, and in order to provide the cell current (both the holding and reading current), the Darlington Configuration played an important role. On the other hand, the word-line amplitude of CMOS memory equals to the full swing, 5V, and no DC current flows in the steady state for a cell.

The function of the CMOS memory is explained as follows; The cell is composed of the cross-coupled CMOS gates, or the flip-flop. The word-line (WL) just opens the gates of pass transistor, connected between the bit-line and cell. Owing to the equal load attached to each bit line, the voltage difference appears between the pair, depending on the conductance of NMOS drivers compositing the flip-flop. Simulated delay time from the word-line selection to the output of the sense amplifier, using the scheme of the 64K CMOS SRAM, is shown in Fig. 3.5. As same

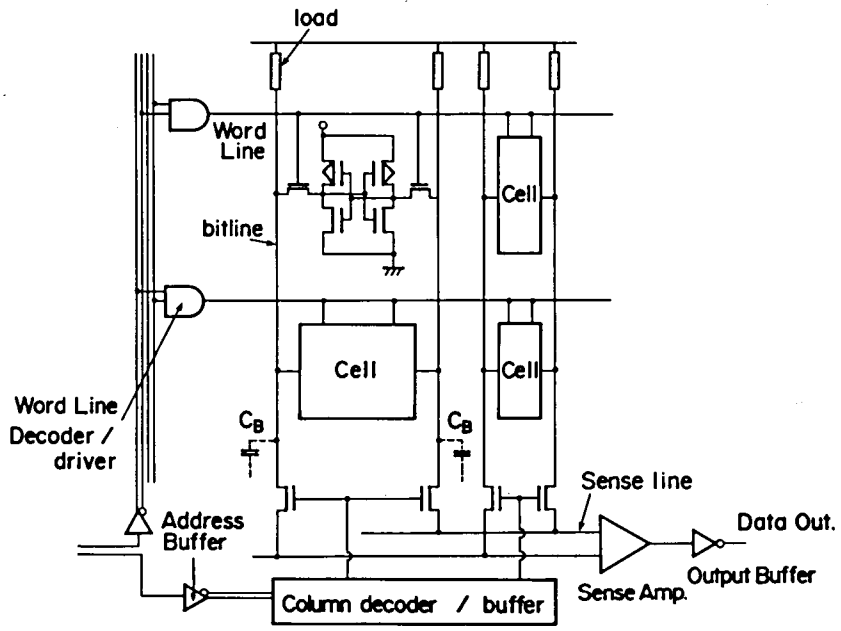


Fig. 3.4 Schematic diagram of CMOS RAM.

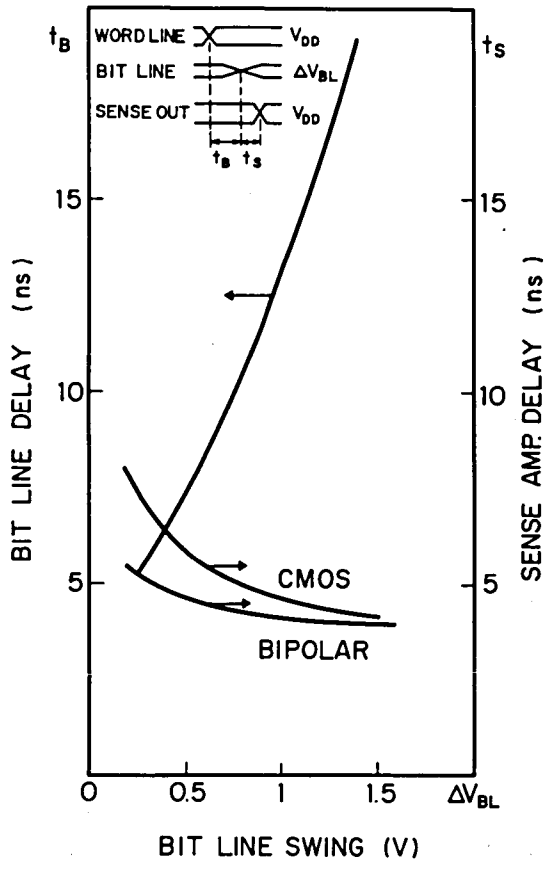


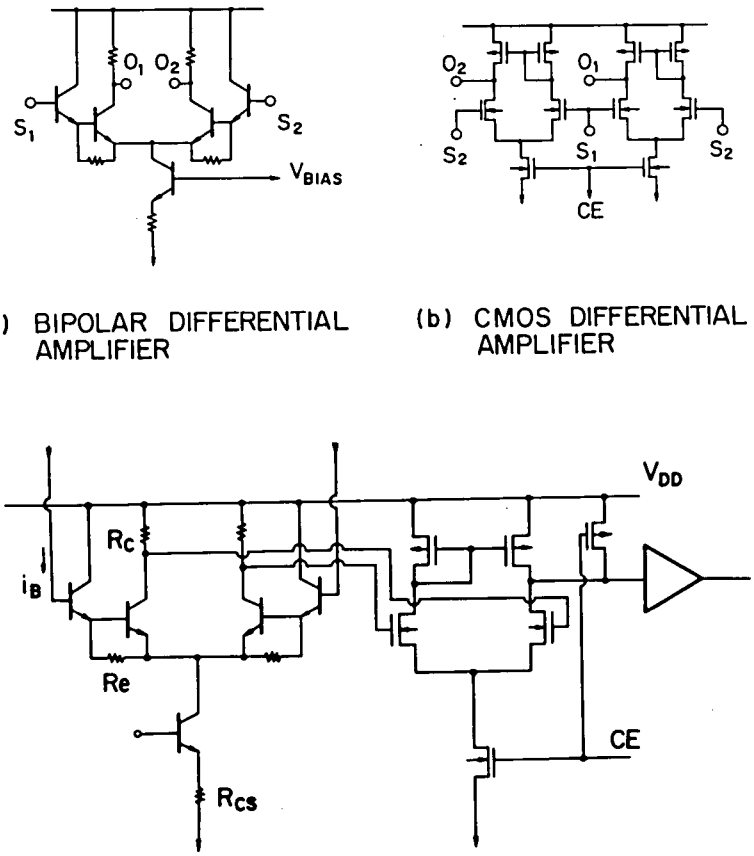
Fig. 3.5 Simulation results of the delay time in bitline and the sense amplifier as a function of bitline swing.

as the ECL memory, the word and bit-line delay is a major part of the total address access time. Both the bit-line delay time, t_B, and the sense amplifier delay time, t_S, is a function of the bit-line voltage swing, ΔV_{BL}, which is determined by the drivability ratio of NMOS driver and pass transistor to the bit-line pull-up load. This result was obtained by keeping the cell drivability constant. It is concluded from the Figure, that t_B decreases almost linearly as ΔV_{BL} decreases. This can be easily understood from the following simple equation,

$$t_B \propto C_B * \Delta V_{BL} / i_c . \qquad (3.1)$$

where, C_B is the parasitic capacitance of each bit line, composed of the drains of unselected pass transistors and bit-line metal to substrate. And it is determined by the cell layout and process parameters. The sinkable cell current, i_c, is also determined by the NMOS dimension, or the cell layout. In order to reduce a cell size, or to put into a pre-determined package, the most advanced design rule is generally applied to the cell. So, C_B, and i_c are the pre-determined parameters. Therefore, it is concluded that the only way to improve t_B is the reduction of ΔV_{BL} by changing the impedance of the bit-line pull-up load. This approach, however, needs high sensitive amplifier, and precise control of the bit line swing, too.

For the actual comparison, two types of differential amplifiers shown in Fig. 3.6, i.e.,(a) an emitter-coupled bipolar differential amplifier and (b) a current mirror loaded CMOS differential amplifier were designed and



(a) BIPOLAR DIFFERENTIAL AMPLIFIER

(b) CMOS DIFFERENTIAL AMPLIFIER

(c) BIPOLAR SENSE AMPLIFIER

Fig. 3.6 Schematic of the bipolar and CMOS differential amplifiers.

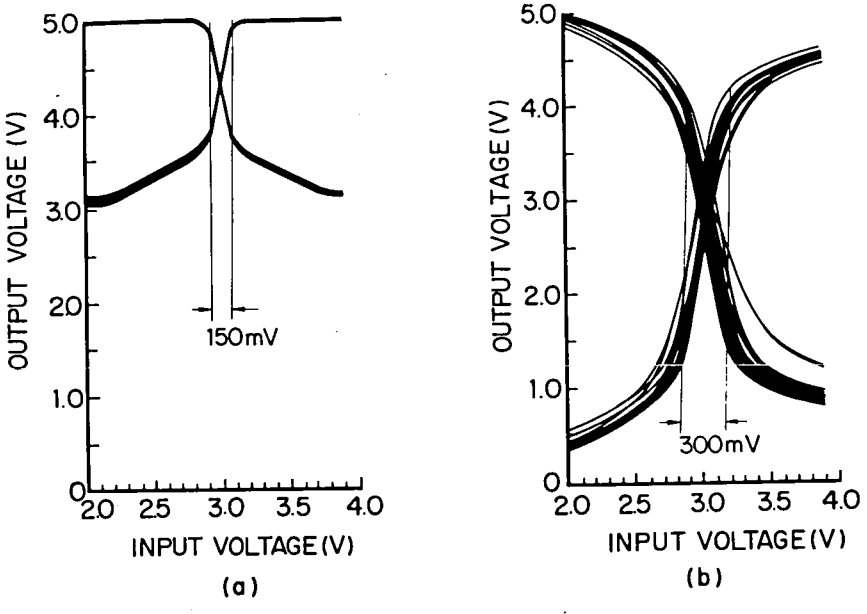


Fig. 3.7 DC characteristics of the (a) bipolar and (b) CMOS differential amplifiers.

evaluated. The Darlington Configuration was incorporated for higher input impedance. The miniaturized cell could provide the base current. The gain, $\partial Vo/\partial Vi$, for CMOS differential amplifier, when one input voltage transits, and become approximately equal to the other input, is calculated as follows by the simple MOS model,

$$\partial Vo/\partial Vi = gm*Z_L , \qquad (3.2)$$

where Z_L is the output load impedance. On the other hand, under the same condition, the gain of the emitter-coupled bipolar amplifier is given as follows,

$$\partial Vo/\partial Vi = 1/4*q/kT*V_L Z_L/(Z_L+R_c)$$
$$= 1/4*q/kT*V_L \quad (Z_L >> R_c) \qquad (3.3)$$

where V_L is the logical swing of the output and R_c is the collector load resistor. The value, kT/q equals about 25mV at room temperature. Equation (3.3) indicates that the gain, or the sensitivity of the bipolar amplifier be independent of Z_L, if $R_c << Z_L$, in contrast with CMOS amplifier. Therefore, the bipolar amplifier potentially has a capability of driving a comparatively lower impedance without sacrificing the gain, which results in its high speed sensing capability in case of small inputs.

However, this bipolar differential amplifier could not play a role of a bit line sense amplifier as it was, because a full swing output necessary to drive a CMOS output buffer could not be obtained by itself. Input and output levels of this amplifier were carefully designed to prevent NPN transistor from the saturation, or to keep the collector-base junction backward biased any time.

Therefore, a second-stage amplifier, followed by the bipolar was required. As shown in Fig. 3.6(c), this was designed by CMOS, which was provided the sufficiently amplified input by the first stage. On the other hand, SPICE circuit simulation suggested that in order to get higher speed sensing for a small amplitude and slow transition input, the CMOS amplifier shown in Fig. 3.6(b), also needs the second stage as a bit line sense amplifier.

Two kinds of bit-line sense amplifiers, both composed of two stages, Bi-CMOS (represented as Bipolar-SA), and CMOS-only (as CMOS-SA) were studied by the circuit simulation. The delay, t_S, as a function of ΔV_{BL} is also shown in Fig. 3.5. It indicates that the Bipolar-SA is faster than CMOS-SA and the former reveals smaller ΔV_{BL} dependence.

The two differential amplifiers used as a first stage shown in Fig. 3.6(a) Bipolar-DA and (b) CMOS-DA, were fabricated on the same chip and evaluated. The both resistance values of R_c and R_e of Bipolar-DA were chosen as 7kΩ, taking internal collector resistance, r_c, into account. The input load current was 2uA, which could easily provided by the minimized MOS cell transistors. The design rule of minimum gate length was 1.5um and 1.8um for NMOS and PMOS for the memory cell, respectively. However, for CMOS-DA, longer gate length of 2um was applied not to lose the matching capability of current mirror transistors. The input and output characteristics of several different devices on a wafer are shown in Fig. 3.7. To get a 1.0V

output, for example, an input signal of 150mV is required for Bipolar-DA, and 300mV for CMOS-DA, taking chip-to-chip device variation into account. In addition, maximum gains of Bipolar-DA and CMOS-DA are 11 and 7, respectively, as predicted from Eqs. (3.2) and (3.3). The power consumption of CMOS-DA was almost 10 times larger than Bipolar-DA, while the occupation area of Bipolar-DA was almost twice larger than CMOS-DA, as shown the photographs in Fig. 3.8. Since the CMOS-DA gave larger chip-to-chip variation, the worst case delay time is implied to be actually larger than that simulated in Fig. 3.6. Bipolar-DA offered larger design tolerance for the parameter variation effect in the memory cells on the access time.

In order to compare the access time of both sense amplifiers, Bipolar-SA and CMOS-SA, test devices having each 8k memory cell array, incorporated with either sense amplifier were fabricated. The standard deviation, σ, of address access time among various chips and the difference of the average access times from Bipolar-SA were obtained as shown in Table 3.1. In the case of Bipolar-SA, ΔV_{BL} dependence ia also shown. From the measurement, the following three results were obtained. First, the memory with Bipolar-SA was 5.6ns faster than CMOS-SA in case of the equal ΔV_{BL}. Second, the access time deviation of Bipolar-SA among various chips was smaller than CMOS-SA. Third, the reduction of ΔV_{BL} was quite effective to achieve high-speed.

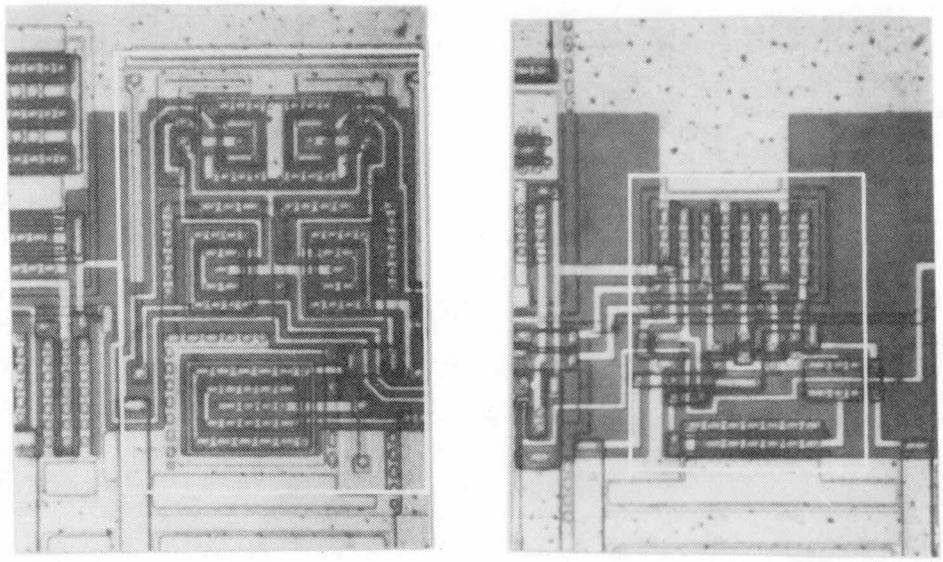


Fig. 3.8 Microphotographs of (a) bipolar and (b) CMOS differential amplifiers.

Tab. 3.1 Speed comparison of the bipolar and CMOS amplifiers.

item	Bipolar SA		CMOS SA
ΔV_{BL} (bit-line)	700mV	860mV	700mV
3σ (access time)	6.3ns	8.7ns	12.3ns
Δt (access time)	0	+8.7ns	+5.6ns

3.4 High Speed 64K SRAM Design

A high-speed 64K Static RAM was designed using the Bi-CMOS circuit technology. This RAM is organized 8k word by 8bits, operating fully asynchronous. The diagram of this RAM is shown in Fig. 3.9. Besides 13 address inputs and 8 data I/O pins, it has three control pins, i.e.,Chip Enable(CE) for making the chip active, Output Enable(OE) for changing the output buffer tri-state to low impedance, and Write Enable(WE) for writing data into memory cells. 8 address inputs select one of 256 word lines, and the rest 5 address select one of 32 bit line pairs by raising the gate of column pass transistor. The RAM is put into 28pin DIP package.

Memeory cells and its peripheral circuits are shown in Fig. 3.10. Bipolar NPN transistors were used to the bit line sense amplifier, voltage regulator, and the bit line clamp, while CMOS circuits were adopted to the other portions such as input buffers, output buffers, second stage amplifiers, row and column decoders and memory cells. The internal wave forms, obtained by the circuit simulator, SPICE, are shown in Fig. 3.11. The labeled numbers in the Figure are corre-eponded to the nodes on Fig. 3.10, each other. The operation of the memory peripheral circuits will be describeded in the following paragraph, using those wave forms.

First, the row address input is provided as ①. It is a worst case input specification of TTL level, swinging between 0.8V to 2.2V. Then, it is converted to the CMOS level, and is amplified by the address buffer. By the decod-er, one of 256 word lines is selected. Although the word

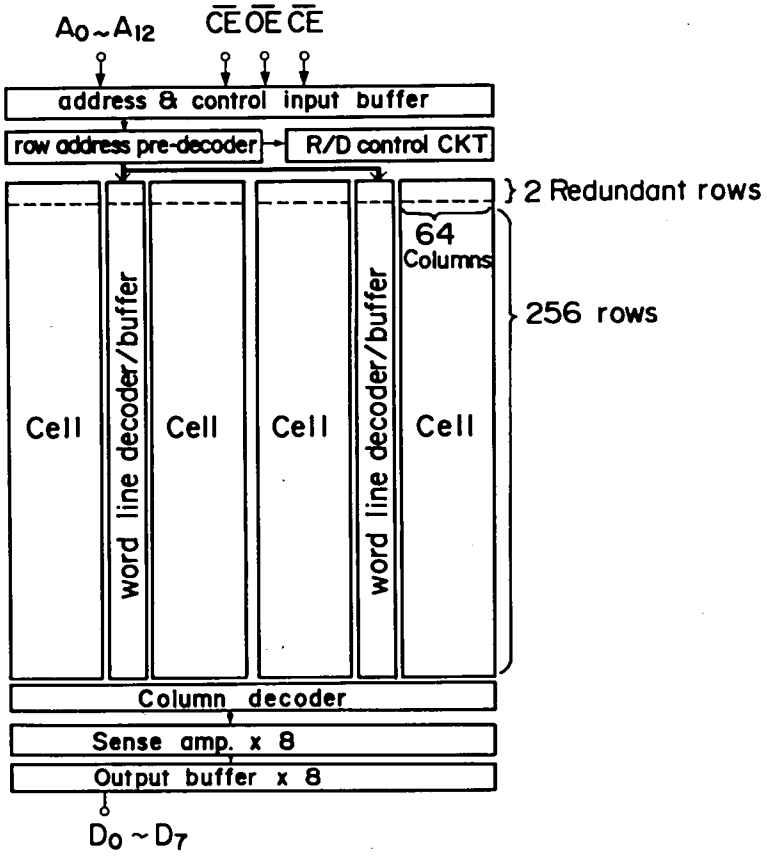


Fig. 3.9 Memory block diagram.

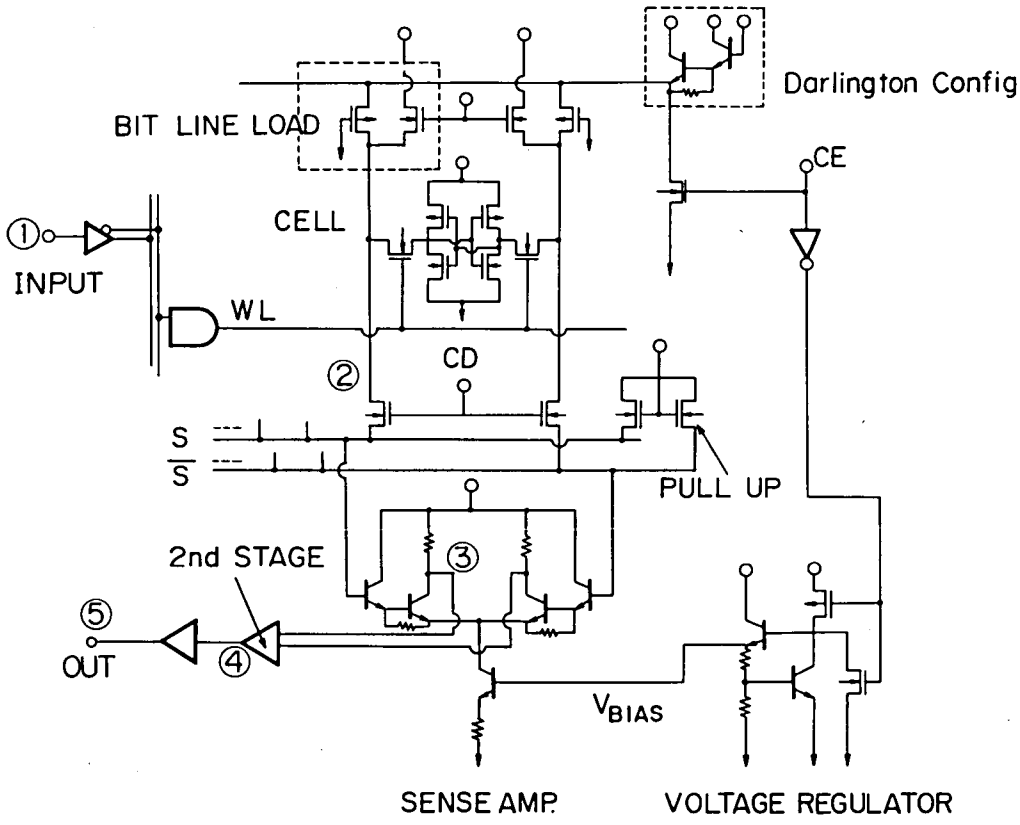


Fig. 3.10 Cell and its peripheral circuitry.

line buffer begins raising a word line, the signal transmits with RC delay to the farest bit line from the buffer, which may determine the worst case delay. Two methods were taken for the word line delay. One is a two decoder architecture, as shown in Fig. 3.9. As the memory cell array is divided by four, one word line should drive only 64 cells. The other was the employment of the Mosi as the gate material, which made the resistivity about one sixth of the conventional N type polysilicon. The resistance and capacitance including the line and gates, was reduced to 4kΩ, and 0.5pF, respectively. Although its RC delay is calculated only 2ns, the bit line pairs, at the farest from the word line buffer, changes very slowly as ②, due to the bit line.

The bit-line amplitude is determined by adjusting the dimensions of bit-line load, and sense-line pull-up transistor, as shown in Fig. 3.12. Here "sense-line" is the input of the sense amplifier, common to 32 bit-lines via the column decoding pass transistors. Normally-on bit line load made by parallel connected NMOS and PMOS defines the bit line high level. The source of PMOS is connected to the common line driven by the NPN Darlington Configurations. By the PMOS shunt transistor, the bit line high level is clamped at VDD $-2V_f$, where V_f is a base-emitter voltage drop, in any supply voltage, if the cell NMOS driver pulls down no current. This clamping voltage is slightly higher than $VDD-V_{th}$, where V_{th} is the threshold voltage of the column transfer gate driven by the column selection, "CD", with the body effect. So, the high level of the sense-line is slight-

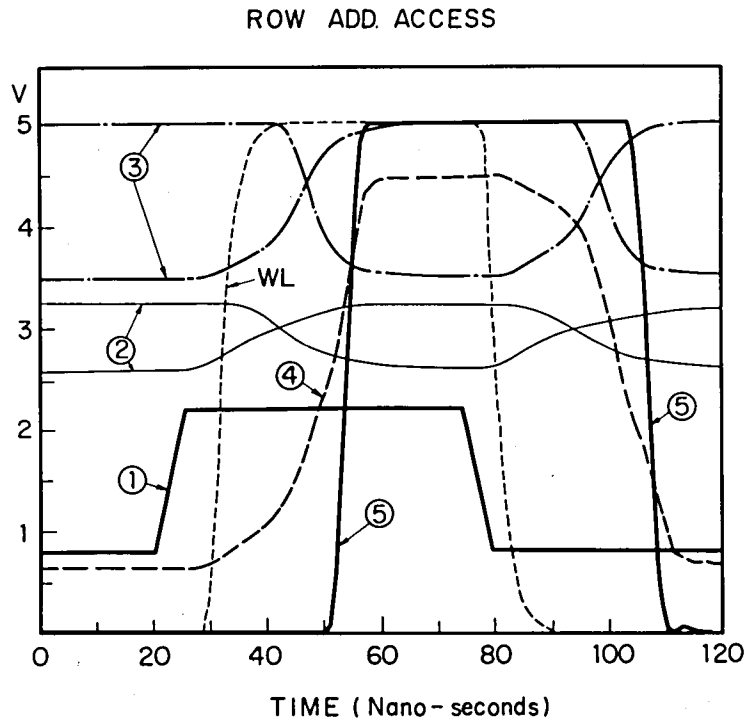


Fig. 3.11 Simulated wave forms of the 64K SRAM. Labels correspond to the circuits' nodes in Fig. 3.10.

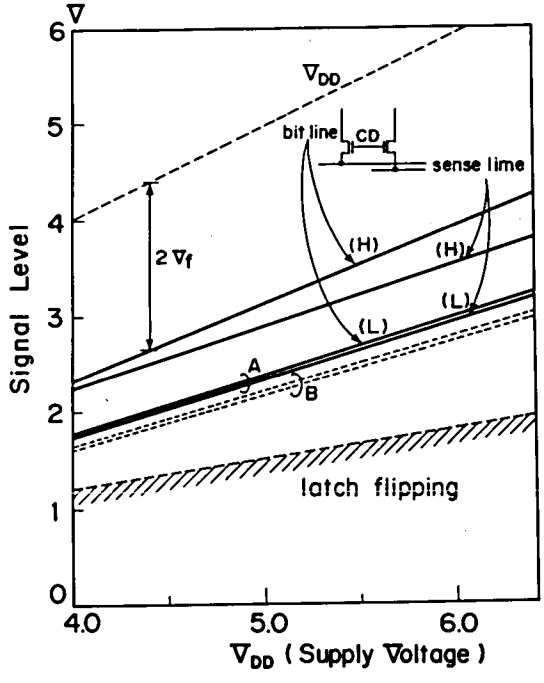


Fig. 3.12 Voltage levels of the bitlines and sense lines as a function of the supply voltage.

ly smaller than the bit-line high. NMOS shares the low bit line current with the bipolar clamping circuits. PMOS is good for shunting the bit line to a higher level, while NMOS provides the sufficient current for a bit-line write recovery. The sense-line pull up was designed to help the sense line recovery, but not to affect the low level setting very much, as shown in Fig. 3.12. By cutting off 25% of the dimension of the bit line load, the low level decreases from A to B, and the bit-line amplitude increases 0.7 to 0.86. Both low levels are larger than the latch flipping voltage. That means even if multi word-lines are selected at some address transitions by some process fluctuations, the stored cell data will never be broken in the read operation.

The small bit-line signal, ②, is first amplified to an appropriate level, or the sufficient level for the second stage CMOS amplifier, with the rapid transition by the bipolar amplifier, ③, and then converted to an almost full swing by the followed CMOS amplifier, ④. As the sense-line high level does not exceed VDD-$2V_f$, the output swing of 1st stage bipolar sense amplifier could be designed as $2V_f$, with a V_f margin to the NPN saturation. The signal is terminated by the output buffer and, the output data is obtained as ⑤. Under the condition of loading 30pF external capacitance, this simulation predicts below a 30ns row access time. As to the column access path, there is no large line transmission delay as the word line. The bit-line levels have been settled before the column pass transistor gets on. So, the column access time would be shorter than the row.

The bias current of the bipolar amplifier is controlled by the internal voltage regulator shown as Fig. 3.13. It generates the voltage of $1.5V_f$, which is determined by the resisters' ratio, to the base of constant current source transistors. Each transistor limits the current flow of the amplifier to 160uA. V_{DD} dependence and the output load dependence of this regulator are shown in Fig. 3.13(a) and (b), respectively. As the regulator provides totally 1.28mA, it is indicated that the regulator can drive the bipolar amplifiers enough, and works stably, even if V_{DD} fluctuates by some noise.

All bipolar devices in these circuits were designed to operate in an active region (the collector-base junction is back-biased) in any mode, in order to avoid latch-up and stored charge delay. In addition,the substrate P region surrounding the bipolar transistor is contacted GND. MOS switches turn off every current path from V_{DD} to GND in the standby mode (CE is low). For example, when CE goes low, the "Vbias" is set to be zero, and the DC current path of both bipolar and CMOS amplifier shown in Fig. 3.6(c) are cut off. The output node of the second stage CMOS amplifier is elevated to V_{DD} to eliminate the idle current of the next CMOS buffer, by PMOS pull-up.

By using six transistor cell structure, only 100nW stand-by power was consumed. The value was caused by the leackage current of the junction. The cell size is 18*20um, where the 1.6um minimum feature size is used for the contact hole, 1,5um for NMOS gate, and 1.8um for PMOS gate

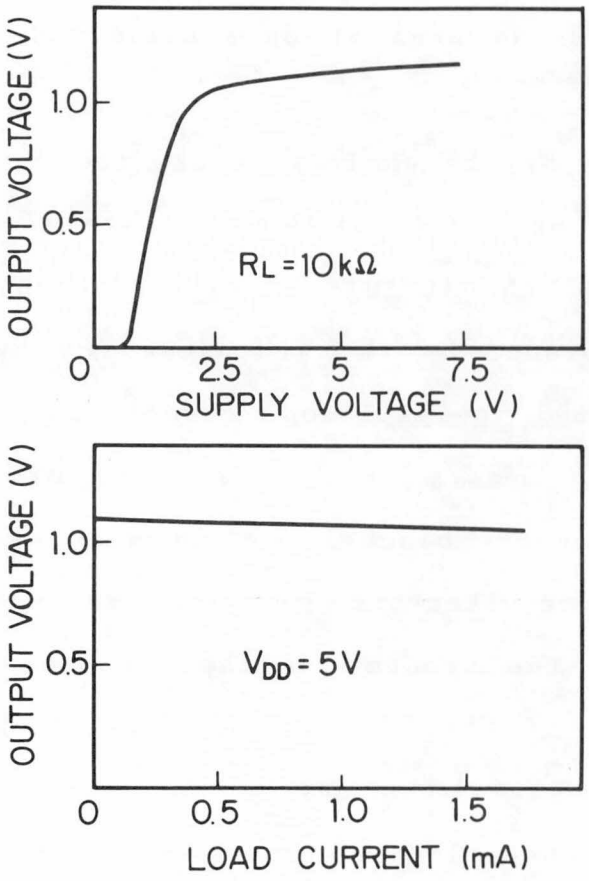


Fig. 3.13 Characteristics of the voltage regulator.

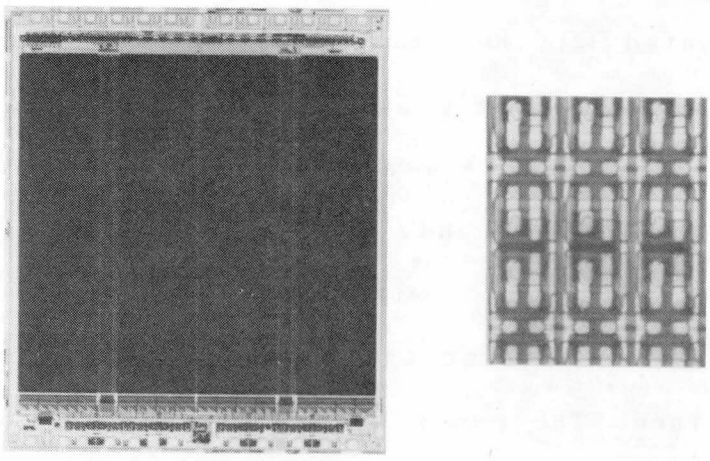
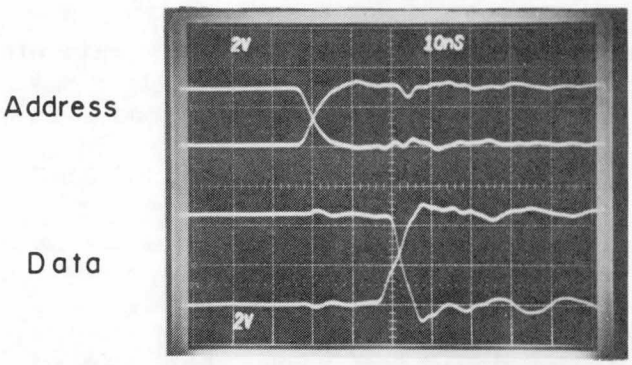


Fig. 3.14 Microphotograph of the (a) whole chip and (b) cells.

length. In the cell, a technique to contact all aluminum, poly, and diffusion area at once minimized the cell size, more.

Photographs of the whole chip and the cell are shown in Fig. 3.14(a) and 3.14(b), respectively. The chip is 5.95*6.84mm, which includes two redundant rows. Fault address is replaced by the redundant row by means of the laser programmed polysilicon fuses. As shown in the photograph, the memory cell arrays are divided into four blocks, and only one block is selected by two block address signals. The architecture reduces the word line delay propotional to the product of the resistor and capacitor, into a quarter of the conventional two block architecture. And it, also, saves the active power dissipation into a half of the conventional one. Because the major current flow in the active mode, occurs from the normally-on bit line loads to the NMOS cell drivers through the pass transistors connected to the selected row, and it increases proportional to the number of bit-lines. Mosi gate material also contributed to the word-line delay reduction. Its sheet resistivity was only $5\Omega/_{\square}$., much lower than $30\Omega/_{\square}$ of N-type polysilicon. The row and column access waveforms are shown in Fig. 3.15(a) and (b), observed under the room temperature, and V_{DD}=5V, with 30pF external capacitance load. A row address access time of 28ns, and a column of 23 ns were obtained. The power dissipation in the active mode was 225mW. The characteristics of this RAM are summarized in Table 3.2.

(a) Column Access



(b) Row Access

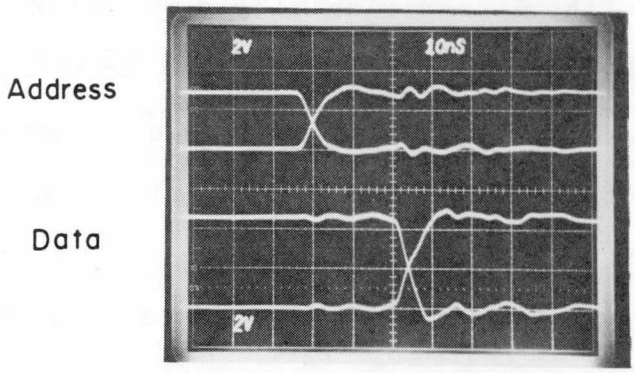


Fig. 3.15 Oscillographs of row and column access time.

Tab. 3.2 Characteristics of high speed CMOS 64k SRAM.

Organization	8kword*8bit
Operation	Fully asynchronous
Technology	N-well CMOS/Bipolar
Cell size	18*20um
Die size	5.95*6.84mm
I/O interface	TTL
Address Access Time	28ns(typical)
Chip Select Access time	28ns(typical)
Active Power	225mW
Standby Power	100nW
Redundancy	2 spare rows
Package	28pin DIP

3.5 Discussions about Bipolar Application in CMOS circuits

Apart from restrictions of the fully compatible process with the current CMOS process, I will introduce N^+ buried layer in this section and make a general disscussion about the application of bipolar devices on the CMOS circuits.

3.5.1 N^+ buried layer

From the design point of view, the use of the N^+ buried layer enhances the performance of Bi-CMOS circuits. For example, the collector load resistance, R_c, was chosen as 7kΩ into bit-line sense amplifiers, taking the internal collector resistance, r_c of 600Ω into account. In a word, in order to obtain the maximum gain, and to avoid the bipolar saturation, r_c should be negligible in comparison with R_c. That is why the operating current was 160uA. However, as indicated in Fig. 3.3, this I_c did not give the maximum f_T value. Furthermore, the product of the internal collector resistance and gate capacitance of the next MOS, R_c*C, delay, calculated as 0.3ns, was not sufficiently small. The one way to increase the performance was to reduce r_c, or to use the N^+ buried layer.

In order to obtain the structure with buried layer, the standard CMOS process needs some steps such as the PEP (Photo Etching Process) for N^+, the diffusion and epitaxial layer growth. Forming deep N^+ region to connect the N^+ buried layer and the collector contact is quite effective. Although this seems to be much complicated in comparison with the current CMOS process, it will be required for the

future high density CMOS LSI strucrure. Because the further shrinkage of the P^+ and N^+ active region distance, will result in latch-up by the parasitic bipolar transistors. Making the well resistance small enhances latch-up immunity.

The β, and f_T dependence on the collector current, I_c, both with and without N^+ buried layer are shown in Fig. 3.16 and 3.17, respectively. As the falling point of β is determined by r_c, about 6 times reduction of r_c was achieved by the buried layer. And the maximum f_T was improved from 2GHz to 2.6GHz, although the I_c required for the maximum f_T was increased from 1mA to 5mA. This shift comes from the increase of collector-substrate junction capacitance, and the larger current is required to charge it. The epitaxial thickness used for the NPN transistor, was 5um. If one takes into the consideration that the thickness of 256bit bipolar ECL RAM was 2um (see Chapter 2), more performane improvement would be possible to make the epitaxial layer thinner, although the impurity profile for PMOS has to be optimized.

3.5.2 Gate delay

A 51 stage CMOS ring oscillator consisted of PMOS and LDD NMOS gates, with the actual aspect ratio of gate width to length, W/L, of 10um/1.3um for PMOS and 10um/1.0um for NMOS, was fabricated, and the gate delay was measured. The delay as a function of supply voltage is shown in Fig. 3.18(a). 110ps propagation delay time with 80uA/stage

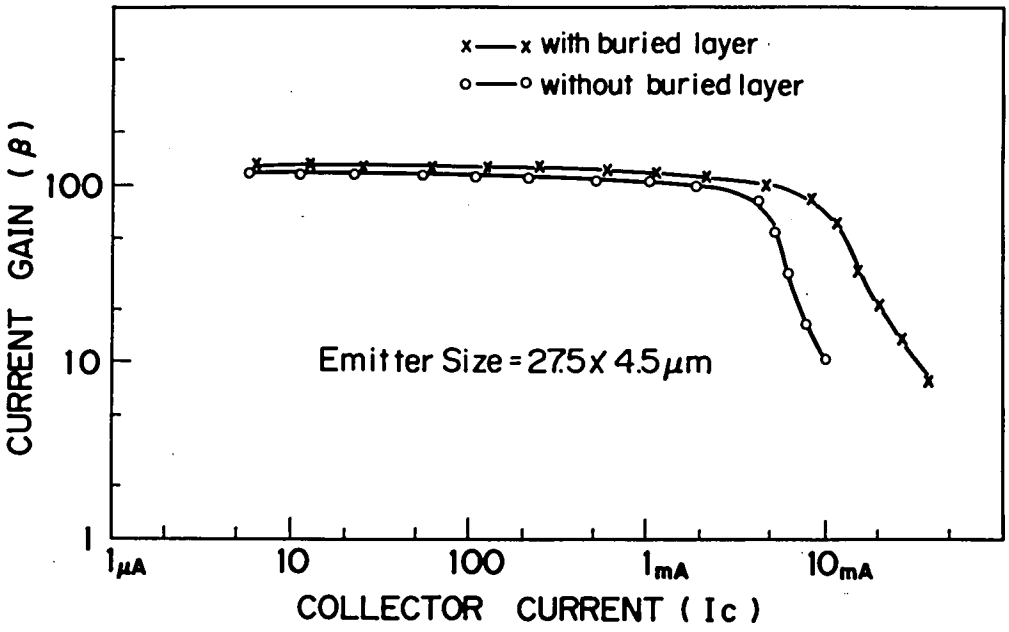


Fig. 3.16 Current gain dependence on the collector current with and without the buried layer.

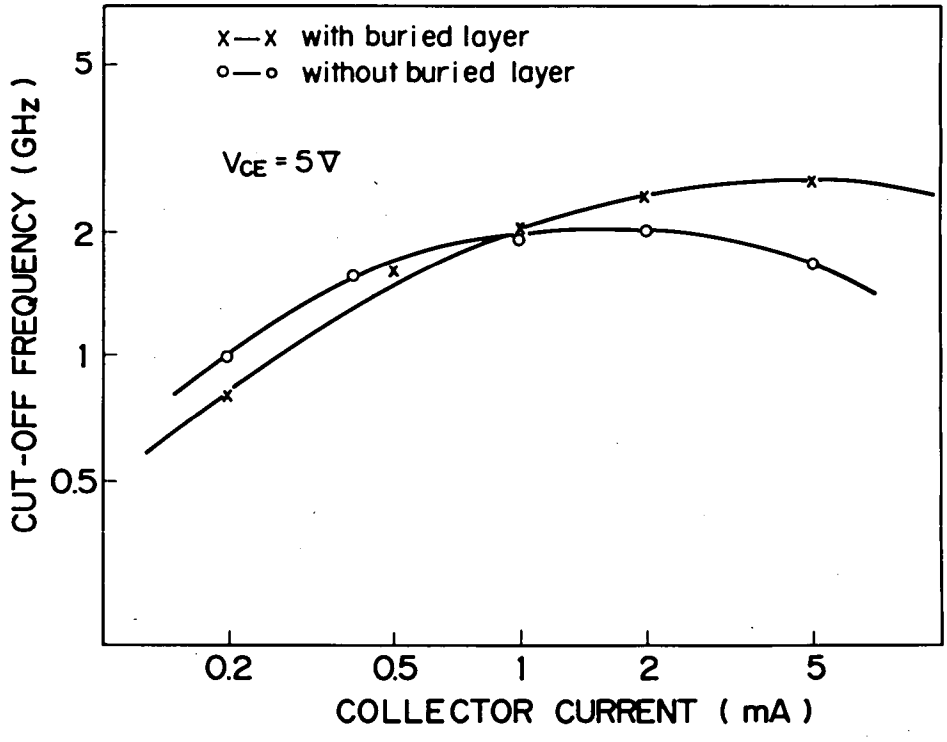


Fig. 3.17 Cut-off frequency dependence on the collector current, with and without the buried layer

supply current was achieved under the 5V voltage supply. At increased 7V, 65ps with 110uA/stage supply current was observed. On the other hand, ECL 19 stage ring oscillator was made by the simple N-well structure and evaluated. The R_c was chosen as 1kΩ. As shown in Fig. 3.18(b), the minimum gate delay was 1.3ns with 690uA. This value could have been improved about one tenth, if we had used the buried layer and had made the optimization of the epitaxial layer thickness and R_c, as well, which should have been 50-100Ω in case of the ECL output buffer. Nonetheless, CMOS could realize the faster gate propagation delay time with smaller current. Furthermore, CMOS has other advantages such as the zero power dissipation in non- transient state and smaller occupation area. Therefore, I can conclude that CMOS is more suitable for the digital parts constructed by the simple logic gates in LSI systems.

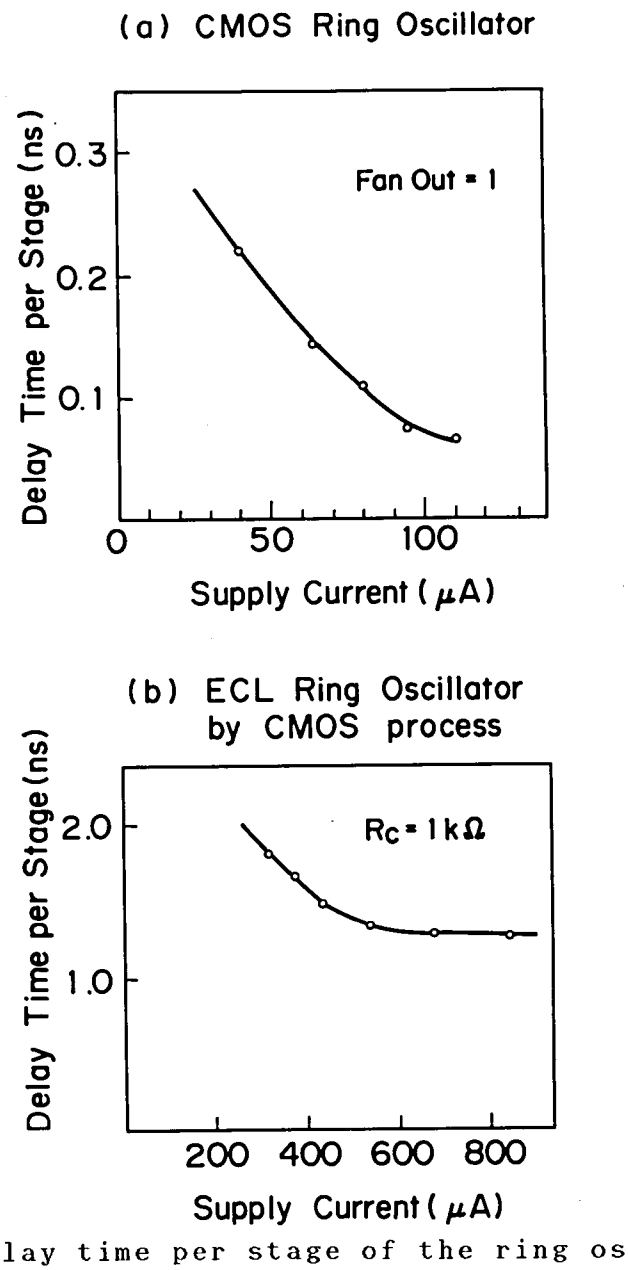


Fig. 3.18 Delay time per stage of the ring oscillators. (a) CMOS 51 stage ring. (b) ECL 18 stage ring.

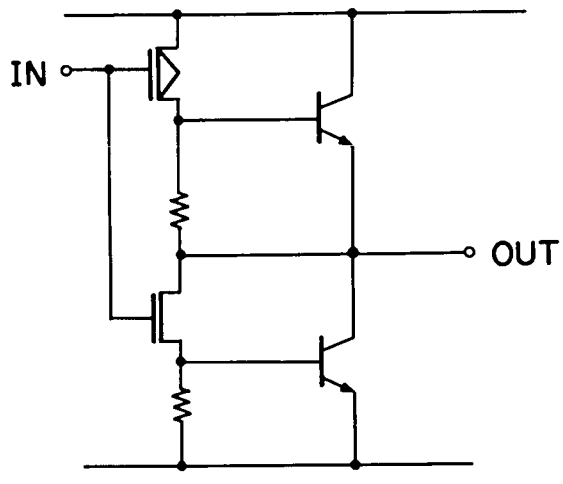


Fig. 3.19 Bi-CMOS buffer circuit.

3.5.3 sensitivity to a small signal.

The conclusion obtained in the previous section should be different for the analog part such as the sense amplifier. In the case of the CMOS amplifier composed of a 1.5um gate length NMOS and a 1.8um PMOS, the ratio of the standard deviation to the average or the gain was 14%, while 9% in bipolar amplifier, as indicated in Fig. 3.7. And the standard deviation, σ of the cross points transiting from high to low or low to high, were distributed from 19mV to 29mV among the wafers in CMOS amplifier, while exact 1mV in bipolar amplifier. The value determines the minimum detectable input signal amplitude. Taking account of 5σ design margin, the minimum signal of CMOS amplifier needs around 100mV, but that of bipolar does 5mV. The overall performance improvement by reducing the signal amplitude has already mentioned in the previous sections. The capability of the small signal detection becomes more important in VLSI, as a scaled device should drive a larger parasitic capacitance in higher speed. The threshold voltage of CMOS is dominated by the controlability of the fabrication process, while that of bipolar is by the built-in potential. The process controlability has been improved at most proportional to the scaling of the device parameters, and it is getting more difficult to exceed it. Therefore, the minimum detectable signal by CMOS amplifier will keep constant, although the gain, which is determined by the gm of MOS transistor may become the same as the bipolar. So, it is

concluded that the bipolar is more suitable for the analog parts such as the amplifier, voltage regulator and level shifter, with the comparatively lower power consumption. The occupation area can be reduced, by improving the minimum design rule, as applied to the CMOS transistors.

3.5.4 Current Driving Capability

In order to compare the current driving capabilities of CMOS and bipolars with and without an N^+ buried layer, a simple output buffer model is introduced. A CMOS output buffer is generally composed of a PMOS transistor for a load device and an NMOS transistor for a driver. Now, the bipolar NPN transistor and the PMOS is compared at first. By the TTL compatibility, the output high voltage, VOH, is specified as 2.4V. Under this condition, the supply current, i, from the NPN emitter follower is calculated as follows;

$$i = (VCC - VOH - VCE(sat))/r_c, \qquad (3.4)$$

where VCC is a supply voltage, nomally 5V, and VCE(sat) is saturation voltage of the NPN transistor, normally 0.2V. The occupied area for the NPN transistor with the buried layer, having r_c of 100Ω, was $30*37um^2$, under 2.0um design rule. So, if 1.0um design rule was applied, the area would be reduced to one fourth of it, or $15*18.5um^2$. The driving capability per area is calculated as $86.5uA/um^2$. On the other hand, the supply current by PMOS follows the MOS triode equation as ;

$$i = gm(VDS(VGS-VTH)-0.5VDS^2), \qquad (3.5)$$

where VDS, the drain source voltage, is now -2.6V, and VGS, the gate source voltage, is -5V. From the measurement of 1.0um PMOS, the current was 0.6mA. As the occupied area was $9.3*7um^2$, the drivable current per area is $9.2uA/um^2$. So, the driving efficiency of the NPN transistor with N^+ buried layer is 9 times larger than that of CMOS. If the buried layer was not applied, the capability would be comparable.

If one wants to use the NPN device as the driver, he has to make the bipolr device operate in the non-saturated region. One method is the use of Schottky barrier Diode, the other is to shunt the collector and base nodes of driver NPN transistor by MOSFET, when the base is elevated to be high. The example circuit is shown in Fig. 3.19, and the performance improvement has been reported in the several papers [25,26]. As the sinkable current generally demands to be larger than the driving current, also N^+ buried layer is necessary for this application.

3.6 Summary

Utilizing a 1.2um MoSi gate Nwell CMOS-bipolar technology, a TTL compatible high-speed 64K static RAM with new CMOS-bipolar circuitry has been developed. Address access time is typically 28ns, with 225mW active power and 100nW standby power. A CMOS six transistor memory cell is used. The cell size is 18*20um, and the chip size is 5.95*6.84mm. The NPN transistors are used in the sense amplifiers, voltage regulators, and level clamping circuits. The bipolar sense amplifiers reduce the detectable bit line swing thus improving the worst-case bit line delay time and the sensing delay time. In order to reduce the large distributed RC delay such as the word line delay, the MoSi layer, which has 5Ω/□ sheet resistivity, is used as gate materials. Two word line decoder/buffer scheme divided the memory cell array into four blocks. It reduced active operating power and the word-line delay to one fourth of the conventional one. The bias current of the sense amplifiers was well controlled by the stable voltage regulators made by the NPN transistors. The bit-line high level was clamped to the appropriate level by bipolar Darlington Configurations. All of the bipolar transistors are carefully designed to operate in the active region with some margin, because of the high speed and latch-up immunity.

The fabrication of the RAM was based on a scaled Nwell CMOS process. Most of the CMOS gates were constructed by 1.2um gate length NMOS and 1.5um PMOS. To avoid the short

channel effect and the breakdown lowering, LDD (Lightly Doped Drain) structure was applied. The transistor guarantees up to 8V break down voltage. Collector isolated NPN transistors and CMOS were fabricated on the same chip without any additional process steps, and without causing any degradation of CMOS characteristics. NMOS channel and NPN base regions were implanted at the same time, and the NMOS and PMOS active region formations are common to the emitter /collector and external base, respectively. The NPN transistor has 2GHz cut-off frequency at 1mA collector current.

The usefulness of bipolar devices in CMOS circuits was demonstrated by the design. In order to achieve high-speed access, the reduction of the bit-line amplitude is quite effective. The bipolar devices are suitable for detecting the small signal. On the other hand, bipolar devices without N^+ buried layer are not faster than the scaled CMOS, when they are applied to the internal logic gates. The propagation delay of CMOS (W/L for NMOS=10um/1.0um, and for PMOS=10um/1.3um) was 110ps with 80uA current dissipation, while the co-fabricated ECL gate was 1.3ns with 690uA. N^+ buried layer will be required for the improvement of both the cut-off frequency and the low impedance drivability. The cut-off frequency of NPN transistors with the buried layer was improved to 2.6GHz. By the DC characteristics of NPN transistor with the buried layer, the drivable current per area was calculated as $9.2uA/um^2$, which was nine times larger than that of CMOS.

64 項欠

Chapter 4

Design of a 256Kbit CMOS EEPROM

4.1 Overview

Recently, electrically erasable and programmable read-only memory (EEPROM) has been developed by the electronic industry due to its in-system reprogramming capability. So far, several kinds of devices up to 64Kbits [27-34] have been intensively demonstrated. They have one standardization of 5V-only operation, which is achieved by the internally generated high voltage taking advantage of extremely low current consumption for their erase/program function. On the other hand, they have revealed many diversities, too. One was demonstrated to have high reliability characteristics, such as over 10^6 endurance with on-chip ECC (Error Correction Circuit), or extended temperature operations. Another aimed at fast access time, 30ns, for example, competitive with bipolar PROM, while the access time of most non-volatile memory including UV-EPROM (Ultra Violet Erasable PROM) is limited to 250ns or so. And there is a synchronous EEPROM like microprocessors operated by some commands. However, in order to apply EEPROM to the non-volatile mass storage for microcomputer peripherals and expand its market, it is necessary to achieve high-density EEPROM with lower cost, in addition to ease of use and high reliability. For this purpose, a miniaturized cell with a simple structure is required.

Utilizing the electron tunneling through thin oxide from the bulk to the floating gate, several types of cells aiming for the future large bit capacity EEPROM were reported [35,36]. However, all of them ended up just the cell proposals, and no high density EEPROM as 256Kbit has been designed. One of the most popular structures is composed of two transistors, that is the select transistor and the floating gate transistor, fabricated in a double polysilicon process. In order to gain larger capacitance coupling between the control gate and the floating gate within the limited stacked area, it is necessary to decrease the interpolysilicon oxide thickness. This would be accompanied by overcoming several fabrication difficulties, for example, the control of the asperity growth, affected by the impurity concentration of polysilicon and the anealing condition [37]. There is another attempt for obtaining effectively thin di-electric film by a triple, oxide-nitride-oxide (ONO), structure. However, it has an inherent problem of charge trapping on the oxide-nitride interface, the mechanism is popular as MNOS device, and the fabrication has not been established yet.

Previously, a single polysilicon cell for 2Kbit EEPROM was reported [38]. However, it occupied 440um^2, which was 2 or 3 times larger than the double polysilicon ones under the same design rule. Because it was designed only for the fabrication process compatibility with the on-chip logic circuits. It is true that the process simplicity would be

attractive, because no special technique is required besides the thin oxide formation on the bulk. The technology has been establishing through smaller density EEPROM fabrication. In order to use the single polysilicon cell for the high density standard memory, the cell size reduction is of vital importance. The larger capacitance coupling between the control gate and the floating gate can be obtained by making the floating gate above the reliable thin oxide on the bulk. For this approach, the cell layout and structure optimization, and detailed investigation of the tunneling effect from the floating gate to the control gate which results in the charge loss, are necessary.

In the high-density EEPROM, there are several hedges to achieve the design goal besides the cell area reduction. One is a data detection technique, because a large parasitic capacitance influences the weak signal generated by the scaled cell, which is a common problem among memory LSI. The second is the erase/programming time reduction, because it increases proportional to the bit density. Utilizing a differential amplifier constructed by a CMOS current mirror circuit [39], or a bipolar one is a way to detect small bit-line signals, as explained in the previous chapter. However, as is used for high speed or high density SRAMs [19], the internally synchronous approach is the another promising method for single ended bit-line cell as EPROM and EEPROM, too. In particular, the scheme has a good compatibility with the page mode programming function, which is effective in reducing the erase/programming time

per byte, virtually.

This chapter describes a 256Kbit 5V only EEPROM fabricated by a single polysilicon and single metal CMOS process. The two kinds of EEPROM cells will be newly proposed in Section 2.2. The layout optimization and the analysis of them will be provided, taking the tunneling effect from the floating gate and the control gate into account. The experimental results will be compared to the analysis. The cell structure and the characteristics actually taken for the 256Kbit EEPROM will be introduced in Section 4.3. The read and erase/program scheme of this memory will be explained in Section 4.4. and 4.5, respectively. The measurement results of the device will be given in Section 4.6.

4.2 Analysis of Single Polysilicon EEPROM Cells

Two kinds of new single polysilicon EEPROM cells (cell A and cell B), composed of the select gate MOS transistor (SGMOS) and the floating gate MOS transistor (FGMOS), are proposed, as illustrated in Fig. 4.1. Cell A is built with a small tunneling area, which is separated from the channel of FGMOS. Cell B is built with a tunneling area coinciding with the channel area, by adapting the thin oxide transistor. Both of their FGs are controlled by the downward N^+ diffused layer through a thin oxide. The reduction of these cell sizes depends on the decrease of erase/program voltage, isolation technology, and design rule scaling. First, an advanced isolation technology SEPOX [40] was applied to minimize the distance between the active regions. Using 1.2um minimum feature size and 1.4um gate length, the cell sizes of A and B are $90um^2$ and $70um^2$, respectively. The erase / program characteristics of these single polysilicon cells will be studied, taking accounts of the effects of the diffused layer control gate(CG).

4.2.1 Erase Operation

The schematic cross section of these cells is shown in Fig. 4.2. The erase conditions are that the control gate voltage, V_C, is raised to an appropriate high voltage, V_{PP}, and that the drain voltage, V_D, equals to zero. Using the symbols on Fig. 4.2, the threshold shift from the initial value, ΔV_{th}, the electric field between the tunneling region and the floating gate, E_1, and that between the floating

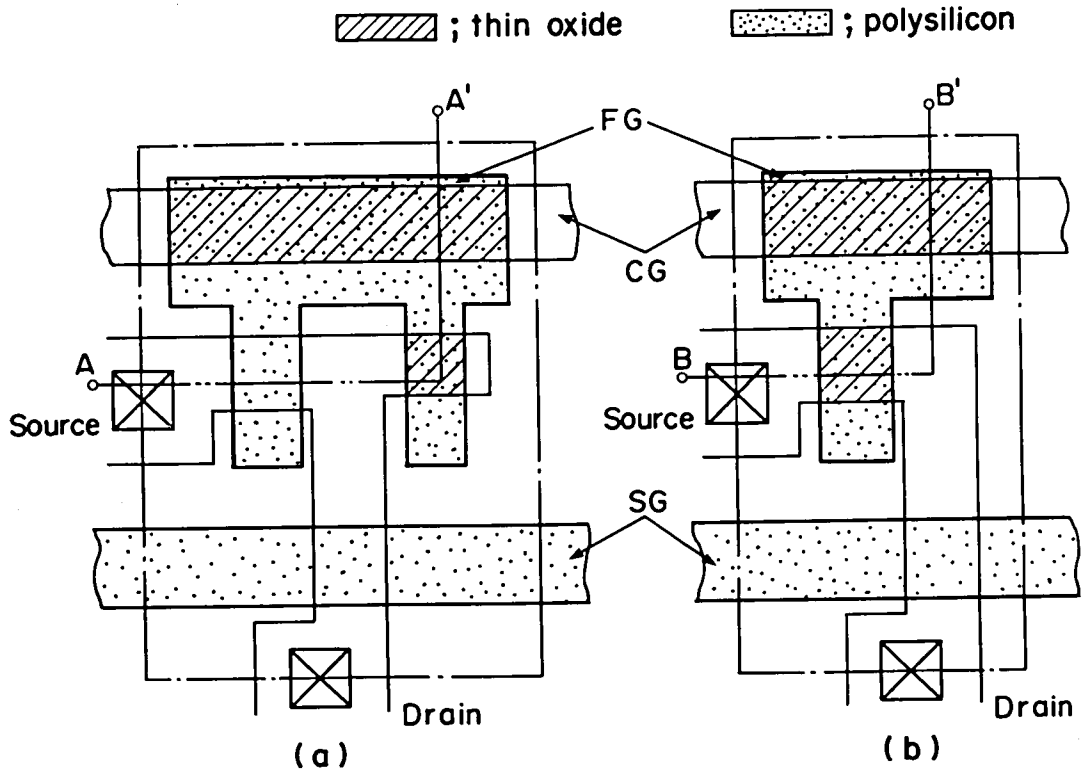
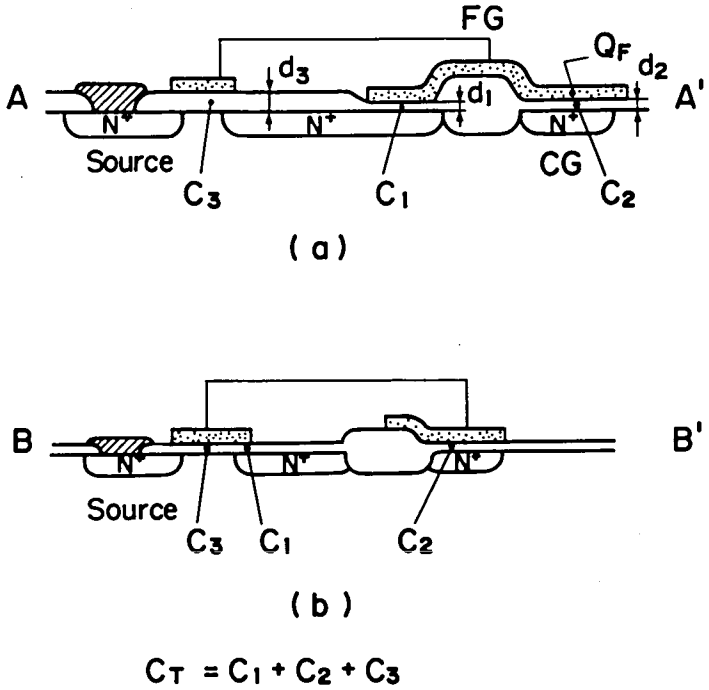


Fig. 4.1 Layout patterns of new single polysilicon EEPROM cells. (a) Cell A, and (b) Cell B.



$$C_T = C_1 + C_2 + C_3$$

Fig. 4.2 Schematic cross section view of the cells. (a) a-a' of the cell A, and (b) b-b' of the cell B.

gate and the control gate, E_2 are written as,

$$\Delta V_{th} = \frac{Q_F}{C_2} \tag{4.1}$$

$$E_1 = \frac{1}{d_1}\,\frac{C_2 V_c - Q_F}{C_T} \tag{4.2}$$

$$E_2 = \frac{1}{d_2}\,\frac{(C_1 + C_3)V_c + Q_F}{C_T} \qquad (C_T = C_1 + C_2 + C_3), \tag{4.3}$$

where, Q_F is the stored charge on the FG, C_1 is the sum of the thin oxide capacitance of the tunneling region and the drain -gate capacitance, C_2 is the control gate capacitance, and C_3 is the gate oxide capacitance, excluding drain-gate capacitance. The value, dQ_F/dt is related to the two components of Fowler-Nordheim current, J(E), as follows,

$$\frac{dQ_F}{dt} = s_1 J(E_1) - s_2 J(E_2)$$

$$= s_1 A E_1^2 exp(-\frac{E_0}{E_1}) - s_2 A E_2^2 exp(-\frac{E_0}{E_2}), \tag{4.4}$$

$$A = \frac{q^3}{8\pi h \phi_B}, \quad E_0 = \frac{-4(2m)^{0.5}\phi_B^{1.5}}{3hq}, \tag{4.5}$$

where all constants have physical meaning and numerical values that are well understood. The parameters, s_1, and s_2, are the areas of the thin oxide on the drain and on the control gate, respectively. The constants, A and E_0 were obtained by the measurement of the thin oxide capacitors as $9.9*10^{-6}A/V^2$ and $2.8*10^8 V/cm$. At the beginning of the erase operation, E_1 is extremely larger than E_2, as this EEPROM

structure has been generally designed as $C_2 >> C_1 + C_3$, and $Q_F = 0$. Then, the current through the CG can be neglected, or $J(E_2) = 0$. Under this conditions, Eqs. (4.1)-(4.4) can be solved analytically, and ΔV_{th} at the erase time, t_e, is given by,

$$\Delta V_{th} = V_c - \frac{d_1 C_T E_0}{C_2}[ln(\frac{s_1 A E_0 t_e}{C_T d_1} + \exp\frac{E_0 C_T d_1}{C_2 V_c})]^{-1} \qquad (4.6)$$

This equation indicates that ΔV_{th} would approach asymptotically to V_c, if the erase time, t_e, infinitely increased. However, the stored charge injected by the tunneling from CG suppresses the FG voltage lowering, as E_2 becomes comparable to E_1. Therefore, ΔV_{th} is considered to be saturated at a certain time, t_S, when $dQ_F/dt = 0$, or,

$$s_1 J(E_1) = s_2 J(E_2) \qquad (4.7)$$

It is to be noted that because of the exponential like nature of the function, J(E), a large relative change in current corresponds to a relatively small change in the electric field. Even though we assume that Eq. (4.6) is satisfied under the condition of $E_1 = E_2$, the relative error is only 2% of the real solution, as long as the ratio, s_2/s_1 is ranging from 1 to 10. It was easily verified by numerical analysis. Therefore, from Eqs. (4.1)-(4.4), the maximum threshold voltage shift, $\Delta V_{th,max}$, is written as,

$$\Delta V_{th,max} = V_c\,[1 - \frac{C_T d_1}{C_2(d_1 + d_2)}] \quad = \frac{V_c}{2}\,(1 - \frac{C_1 + C_3}{C_2}) \qquad (4.8)$$

$$(if\ d_1 = d_2)$$

The time, t_S, when ΔV_{th} equals to ΔV_{thmax}, is calculated from Eq. (4.6) to be,

$$t_s = \frac{C_T d_1}{s_1 A E_0} \exp\left(\frac{E_0(d_1 + d_2)}{V_c}\right) X \left[1 - \exp\left[-\frac{E_0}{C_2 V_c}(C_2 d_2 - C_1 d_1 - C_3 d_1)\right]\right]. \quad (4.9)$$

Taking accounts of the following necessary conditions for the tunneling,

$$\frac{C_1 + C_3}{C_2} \ll 1, \quad E_0 \frac{d_1}{V_c} \gg 1 \quad (4.10)$$

The second term of Eq. (4.9) is assumed to be small. Consequently, t_S is approximately given by,

$$t_s = \frac{C_T d_1}{s_1 A E_0} \exp\left[\frac{E_0(d_1 + d_2)}{V_c}\right] = \frac{C_T d_1}{s_1 A E_0} \exp\left[\frac{2E_0 d_1}{V_c}\right] \quad (d_1 = d_2) \quad (4.11)$$

From Eqs. (4.8) - (4.10), we can summarize the useful characteristics of single polysilicon EEPROM cells. ΔV_{thmax} is determined by V_c, and $C_2/(C_1+C_3)$, and it is almost independent of the thin oxide thickness. Larger ratio of $C_2/(C_1+C_3)$ gives larger maximum threshold shifts. The erase time depends on the ratio, d_1/V_c, exponentially. It is necessary to derease this ratio in order to reduce t_S. After the erase time exceeds t_S, ΔV_{th}, and FG voltage keep constant, which will prevent these EEPROM cells from over erasing. So, the control of the threshold shift is easier. In addition, if the control and the drain thin oxide regions are formed simultaneously, the ratio, $C_2/(C_1+C_3)$, is determined by the cell layout and is hardly affected by the process fluctuation.

The threshold shifts in the erase operation as a function of the erase time were simulated under the condition of d_1=83Å, $C_2/(C_1+C_3)$=5, t_c=2ms, for several values of the control gate voltages, as shown in Fig. 4.3. The calculation results in case of neglecting the tunneling effects between the CG and FG, the case of the conventional double polysilicon cells, are also shown in Fig. 4.3. ΔV_{th} saturates at 1ms , under a 15V applied voltage, which could be predicted from Eq. (4.11).

Consequently, even if it was possible for the double polysilicon cells to have the same capacitance ratio in a reasonable cell area, single polysilicon EEPROM would have no drawbacks as far as the erase time specification of around 1ms is taken. The erase time has a tendency to be shorter, as more bytes are integrated on a chip. As indicated the cell layout in Fig. 4.1, one of the most dominant factors for the cell size reduction is the distance between the CG and FGMOS. The threshold voltage of the field transistor must be larger than the maximum applied voltage to FG. As it is limited by the equilibrium condition of the both side tunneling current as indicated in Eq. (4.7), the value to be guaranteed is, at most, a half of the applied voltage plus some margin. Therefore, the distance between the active regions is not so large, and does not sacrifice the cell size. So, it is concluded that these single polysilicon layout in Fig. 4.1 are suitable for the high density EEPROM cell.

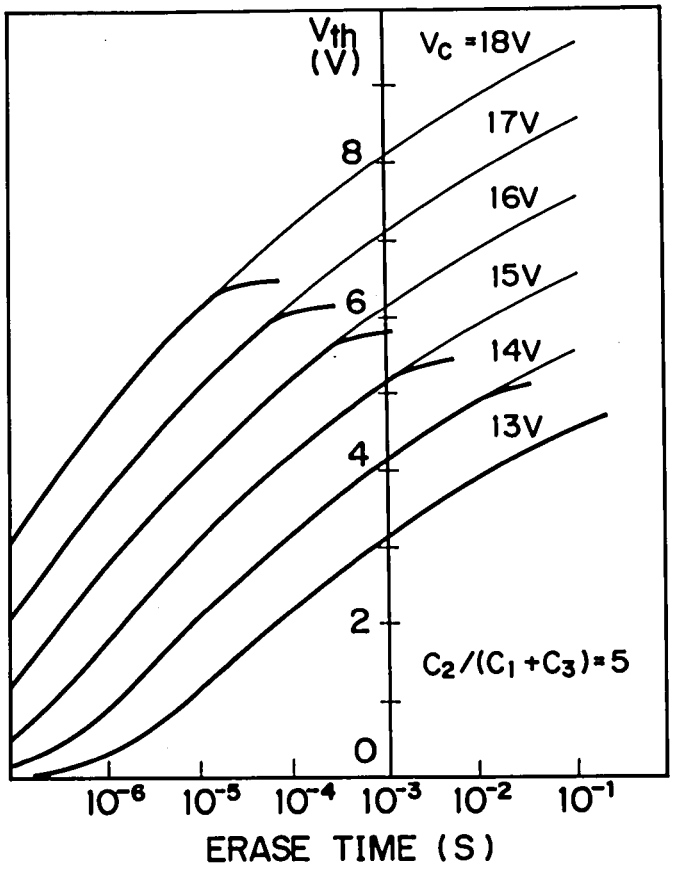


Fig. 4.3 Simulation results of the threshold shift in the erase operation, in case of neglecting the tunneling effect through the diffused control gate (thin line), or including the effect (fat line). $d_1 = d_2 = 83$Å.

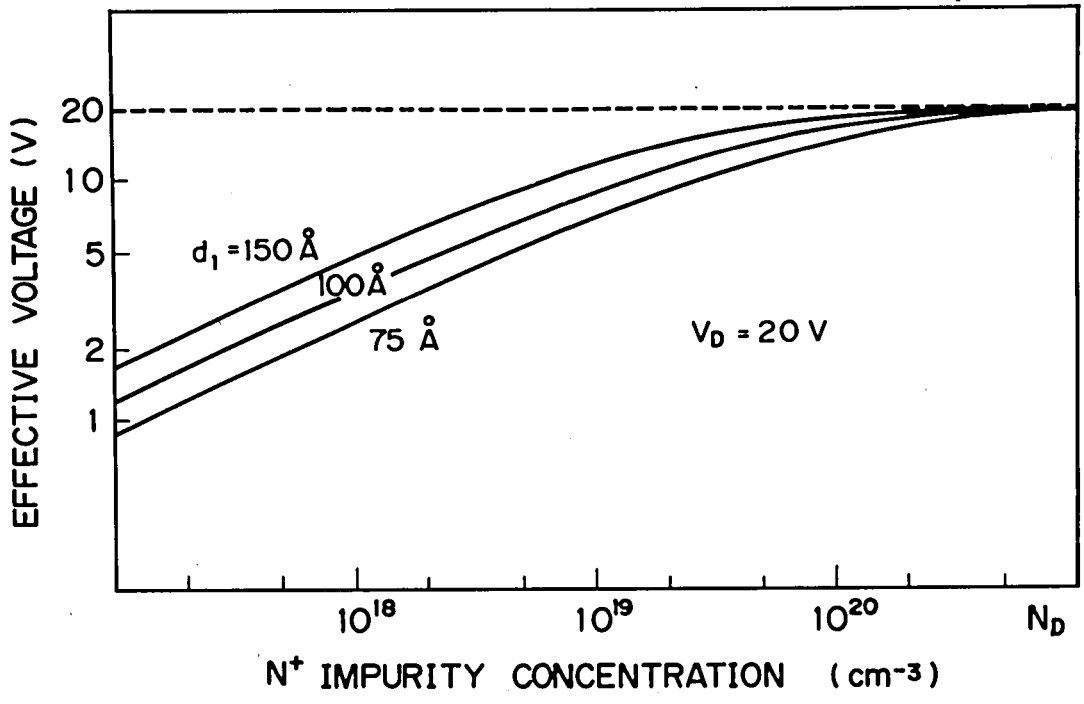


Fig. 4.4 Effective voltages applied to the thin oxide as a function of the impurity concentration, N_D, below the thin oxide, when the depletion region is built.

4.2.2 Program Operation

The program chracteristics can be obtained by the same procedure as the erase operation, except for counting the depletion region effect. Under the conditions of $V_C=0$, and $V_D=V_{PP}$, the FGMOS is driven into an off-state. So, the depletion region, which shares the applied voltage with the thin oxide, is built and quite dominates the tunneling current flow. The simple depletion region approximation gives the effective applied voltage to the thin oxide, as follows,

$$V_{D,eff} = \frac{-\beta + \sqrt{\beta^2 + 4\beta V_D}}{2} \qquad \beta = \frac{qK_s d_1^2 N_D}{2K_0^2 \varepsilon_0} \tag{4.12}$$

Figure 4.4 shows the effective voltage applied to the thin oxide as a funcion of the impurity concentration below the thin oxide. It indicates that even if the N-type impurity concentration around $10^{20}cm^{-3}$ is used, which is as same as that of the NMOSFET's source and drain, the voltage drop across the depletion region is over 10%.

In the program operation, E_1 and E_2 are given by,

$$E_1 = \frac{(C_2 + C_3)V_{D,eff} + Q_F}{C_T d_1} \qquad E_2 = \frac{C_1 V_{D,eff} - Q_F}{C_T d_2}. \tag{4.13}$$

As the tunneling effect through the CG is negligible initially, the threshold voltage shifts, ΔV_{th} at the programming time, t_P, is given by,

$$\Delta V_{th} = \frac{C_T d_1 E_0}{C_2} \left[ln\left(\frac{s_1 A E_0 t_p}{C_T d_1} + \exp\frac{E_0 C_T d_1}{(C_2 + C_3)V_{D,eff} + C_2 V_{Ti}}\right)\right]^{-1} - \left(1 + \frac{C_3}{C_2}\right)V_{D,eff}, \tag{4.14}$$

hence, V_{Ti} is the initial threshold shift at the beginning of the programming, or the threshold voltage after erasing. The minimum threshold shifts, $\Delta V_{th,min}$ and the saturation time, t_S, are approximately obtained by the condition of $E_1 = E_2$, as follows,

$$\Delta V_{th,min} = \frac{V_{D,eff}}{2}\left(\frac{C_1 - C_3}{C_2} - 1\right) \qquad t_s = \frac{C_T d_1}{s_1 A E_0} \exp\frac{2E_0 d_1}{V_{D,eff}} \qquad (4.15)$$

$$(d_1 = d_2).$$

The above equations indicate that in order to increase the absolute value of $\Delta V_{th,min}$, and to make t_S short, it is necessary to increase the ratio of $C_2/(C_1-C_3)$, and V_{Deff}/E_0, which are similar to the erase operation. The value, V_{Ti}, affects neither to $\Delta V_{th,min}$ nor to t_S.

Using the Eqs.(4.6) and (4.15), the erase and the program time can be compared. The critical erase and program time, $t_{e\,crit}$, and $t_{p\,crit}$, are defined as the time when the variable, t becomes large enough and influences ΔV_{th} changings. These are equivalent that the first term of a dominater becomes the same order of the second one. Then,

$$\ln t_{e,crit} \approx \alpha + \frac{E_0 C_T d_1}{C_2 V_c} \qquad \ln t_{p,crit} \approx \alpha + \frac{E_0 C_T d_1}{(C_2 + C_3)V_{D,eff} + C_2 V_{Ti}}. \qquad (4.16)$$

where the common term, α is independent of V_c or V_{Deff}. So, if we assume that $V_c = V_{Deff}$, $t_{p\,crit}$ is smaller than $t_{e\,crit}$, because of the terms of the off-state gate capacitance, C_3 and V_{Ti}. Therefore, by adjusting the impurity concentration below the thin oxide, one can get the same threshold shifts in erase and programming within the same operation time, which

might provide the wide threshold window resulting in the easy data detection.

4.2.3 Transient Analysis

In order to apply the single polysilicon cell to the real memory, the rise and fall time effects on the cell characteristics should be studied. By the transient calculation of the erase and program operation, the wave forms of the floating gate voltage, ΔV_{th}, and the tunneling current through the thin oxide are obtained as shown in Fig. 4.5. This simulation includes all parasitic effects caused by the tunneling through the control gate and the built-in depletion region. Naturally the tunnel current has a peak at the rising edge of the drain and control gate voltages. The ΔV_{th}, and the peak current in both modes as a function of the rise time is shown in Fig. 4.6. The peak current gets reduced as the rise time decreases, while ΔV_{th} is almost independent of the rise time. The rapid change in voltage results in the higher tunnel current, which may cause the permanent damage to the thin oxide, and then the rapid change can not do good for the wider ΔV_{th} window, besides the shorter erase and program time. On the other hand, the fall time did not affect to the characteristics. Then, the simulation indicates that ΔV_{th} is independent of the rise and fall time. So, the previous analysis is valid for the investigation of the characteristics for the single polysilicon EEPROM cell, even though it neglected the rise and fall time effects.

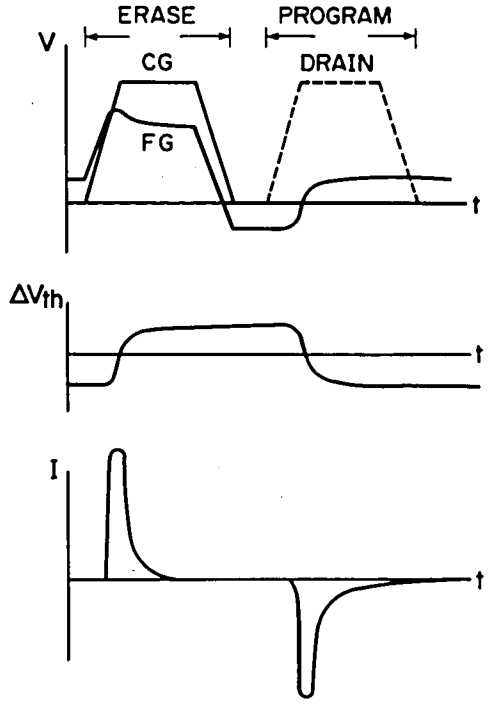


Fig. 4.5 Waveforms of the floating gate, threshold shift, and the tunneling current in both the erase and program operation.

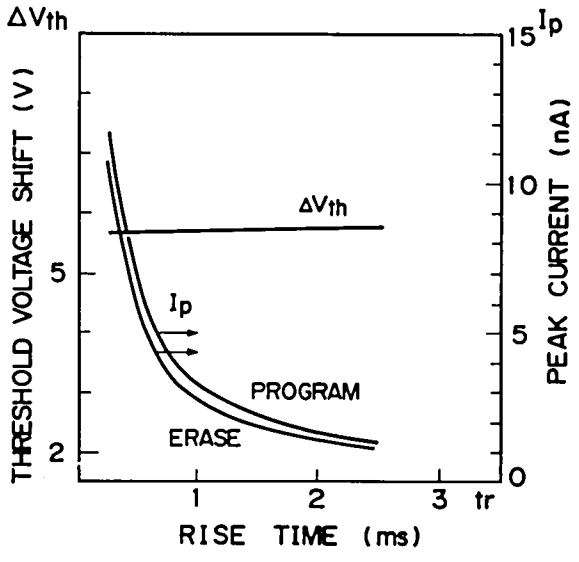


Fig. 4.6 Threshold voltage shifts and peak tunneling current as a function of the rise time of the control gate, and the drain voltages. $d_1 = d_2 = 83$Å, $V_C = V_D = 15$V, and $t_e = t_p = 2$ms, excluding the rise and fall time.

4.2.4 Experimental results and Discussions

In order to investigate the characteristics of the erase and program operation, two types of single polysilicon EEPROM cells, which are already shown in Fig.4.1, type A, and type B were fabricated. Cell A has a separated tunneling electrode from the FGMOS drain. The drain concentration can be chosen to be lower than that below the thin oxide (tunneling region) independently. By means of depletion region effects, the breakdown voltage of the FGMOS can be designed to be high by reducing the drain-gate electric field, allowing tunneling current flow through the thin oxide. Figure 4.7 shows the erase/program characteristics of cell A as a function of the applied voltages, V_C, and V_D. The threshold shifts agree with the simulated results, which are illustrated by solid lines on Fig. 4.7. As predicted from Eq. (4.9), the threshold window becomes large, as both applied voltages and the capacitance ratio increase. Even if the minimum feature size was applied to the diffused control gate area, the ratio, $C_2/(C_1+C_3)$ would be about 3. Utilizing this cell structure, the ratios of 5 and 10 can be achieved with 4% and 14% larger cell areas than the cell with ratio 3, respectively. So, the larger cell ratio can be chosen without sacrificing the larger area in contrast with the conventional double polysilicon cells. To get over 5V threshold window, for example, the conditions as d_1=83Å, $C_2/(C_1+C_3)$=5, t_e, t_P=2ms, V_C, V_D=15V are required in cell A.

Cell B has the tunneling electrode area which is merged

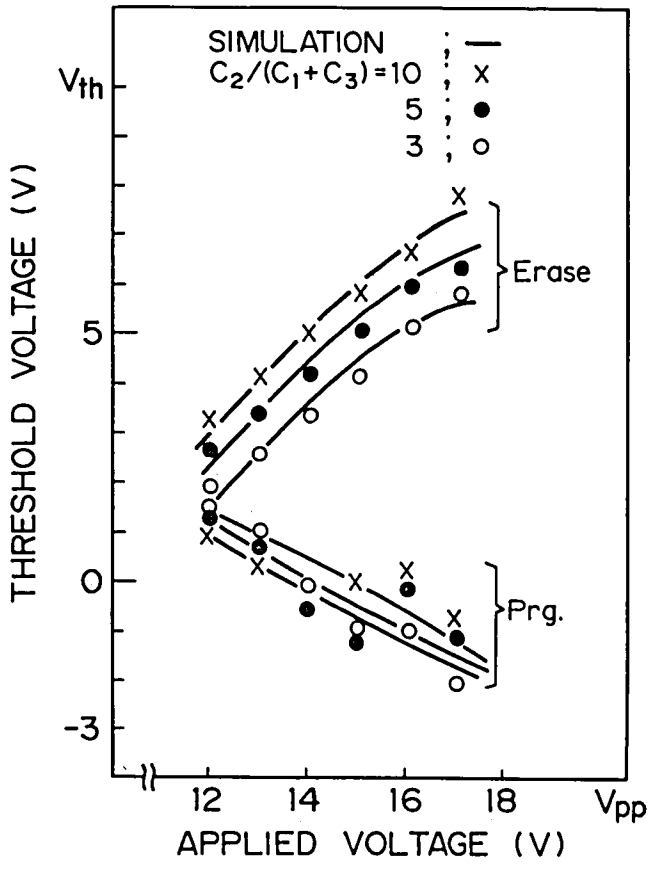


Fig. 4.7 Erase/Program characteristics of cell A as a function of the applied voltage.

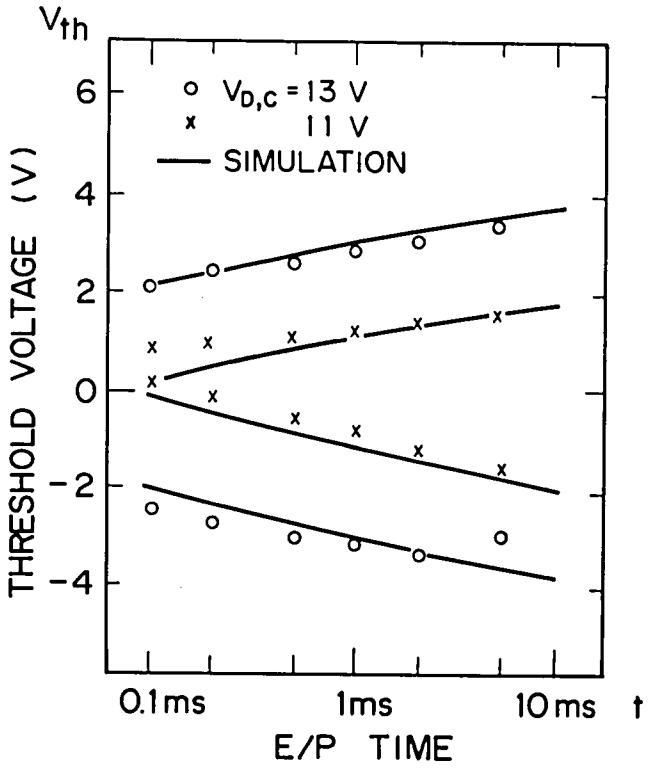


Fig. 4.8 Erase/Program characteristics of Cell B as a function of the erase and program time.

into the FGMOS drain. So, it realizes smaller cell size than cell A. Because of the thin oxide transistor structure, its conductance was observed almost twice as larger than that of cell A, in the same FGMOS dimension, which possibly induces the higher-speed sensing capability. In addition, as the FGMOS is off in the program, the ratio $C_2/(C_1+C_3)$ becomes larger than the erase operation, which makes ΔV_{th} larger than cell A in program. The threshold shifts of cell B as a function of the erase/program time is shown in Fig. 4.8. To get 5V threshold window, about 13V is required under the same device parameters as cell A in Fig. 4.7. And the simulated results also agree with the experimental characteristics.

On the other hand, however, cell B has a disadvantage, caused by the drain structure of the thin oxide FGMOS. As explained above, in order to raise the tunnel current from the drain to the FG in the program, the highly doped drain is required. Highly concentrated electric field applied to the PN junction below the gate polysilicon causes the drain avalanche breakdown voltage lowering. Therefore, this leakage current possibly becomes comparable to the tunneling current. This was verified by the measurements of the thin oxide transistor as shown in Fig. 4.9. As the drain voltage is raised under the conditions of both the source and gate biases are zero, the drain leakage current is divided into two components. One is the gate current resulted from the thin oxide tunneling effect, and another is the substrate current caused by the avalanche breakdown.

The drain impurity concentration and gate oxide thickness were $10^{20}cm^{-3}$ and 83Å, respectively, the same conditions as Fig. 4.8. It is shown that the substrate current is always larger than the gate current with this device parameters. It means that the high voltage generator for the programming current has to supply the substrate current, also. In use of cell B, the internal high voltage generator should provide 1000 times larger current than in use of cell A in order to achieve the single 5V programming, which results in tremendous area penalty for larger charge-pump capacitance as the maximum usable oscillator frequency in IC is limited. However, it is to be noted that in this comparison, the drain profile and the oxide thickness are assumed to be same. By relaxing the electric field between the substrate and the drain, which is accomplished by optimizing the drain impurity profile, the substrate current can be reduced. On the other hand, the gate current increases by making the oxide thin. So, the difference, 3 order of magnitude, may not be kept in future technology.

There is another problem which should be taken into account in use of cell B. Both the thin gate oxide and highly doped drain structure as are used for cell B, are liable to generate the hot electron, when both the drain and gate are biased to high voltage. The mechanism is utilized for the UV-EPROM programming. So, the drain and the CG voltage in the read operation should be designed to be lower enough so as to avoid the read disturbance (soft write).

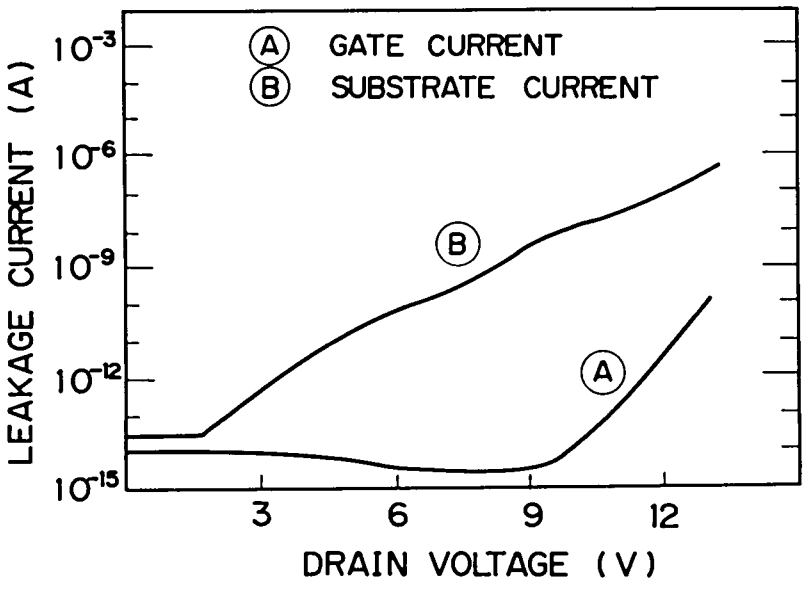


Fig. 4.9 Gate and substrate current of the thin oxide transistor. The drain profile was fabricated similar to the floating gate transistor of cell B.

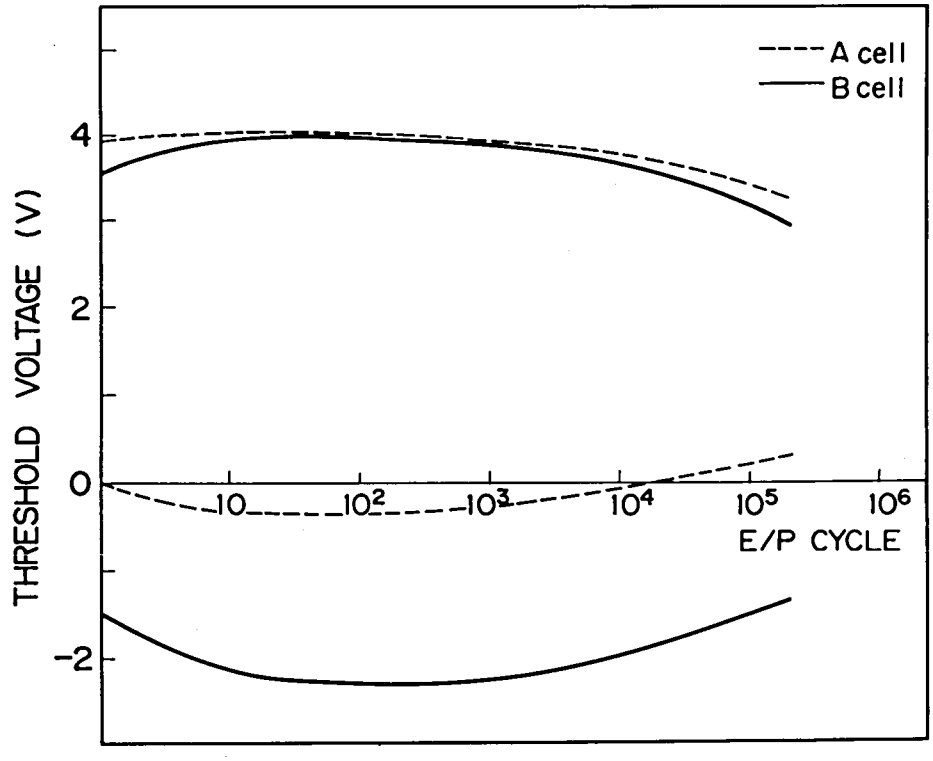


Fig. 4.10 Endurance characteristics of cell A (broken line) and cell B (solid line).

Figure 4.10 shows the endurance characteristics of both types of cells. It indicates that both satisfy 10^5cycle specification, which has been claimed by the conventional double polysilicon cells.

Table 4.1 shows the comparison between cell A and B. There is much trade-off between cell A and B, or the trade-off between the fabrication and design difficulties. The choice depends on the available technology in the LSI generation.

Tab. 4.1 Comparison of device performance between cell A, and cell B.

item	cell A	cell B	better
Cell size	$90\mu m^2$	$70\mu m^2$	B
No. of transistors	2	2	same
Read Current (PL=2.5V, WL=5V)	$100\mu A$	$200\mu A$	B
ΔVth (Vpp=13V, t=1ms)	4.2V	6.0V	B
Endurance	$>10^5$	$>10^5$	same
Programming Current	0.1nA	$0.1\mu A$	A
Soft-Write Immunity	high	low	A

4.3 Cell Structure used for the 256Kbit EEPROM

Based on the study in the section 2, type A was decided to be suitable for the 256Kbit EEPROM. The cell, which is shown again in Fig. 4.11, was named DIFLOX (DIFfused Layer controlled floating gate type cell with thin OXide). The equivalent circuit is shown in the same figure. The capacitance ratio, $C_2/(C_1+C_2+C_3)$ is mainly determined by the thin oxide area ratio, as both were formed simultaneously, and is designed as 0.83. The equivalent circuit of the DIFLOX cell array configurations is shown in Fig. 4.12. The program line, PL, is connected to the diffused control gate via pass transistors, which are controlled by the word line. It is to be noted that the N^+ region of the control gate is common to only one byte (eight cells). The N^+ implantation dose for this region must be chosen carefully. A lower concentration suppresses the effective applied voltage to the thin oxide in the negative gate bias condition, while a higher concentration lowers the barrier height of the thin oxide [11]. Under the optimized concentration applied to DIFLOX, the resistance and the junction capacitance of the N^+ control-gate region were measured to be 20kΩ and 14fF, respectively. The RC time constant is calculated to be lower than 1ns. So, the delay caused by this diffused layer is not a dominant factor on the access time.

To reduce the cell area of DIFLOX, the isolation between active regions is one of the important technologies. A new isolation technology, SEPOX [40], was

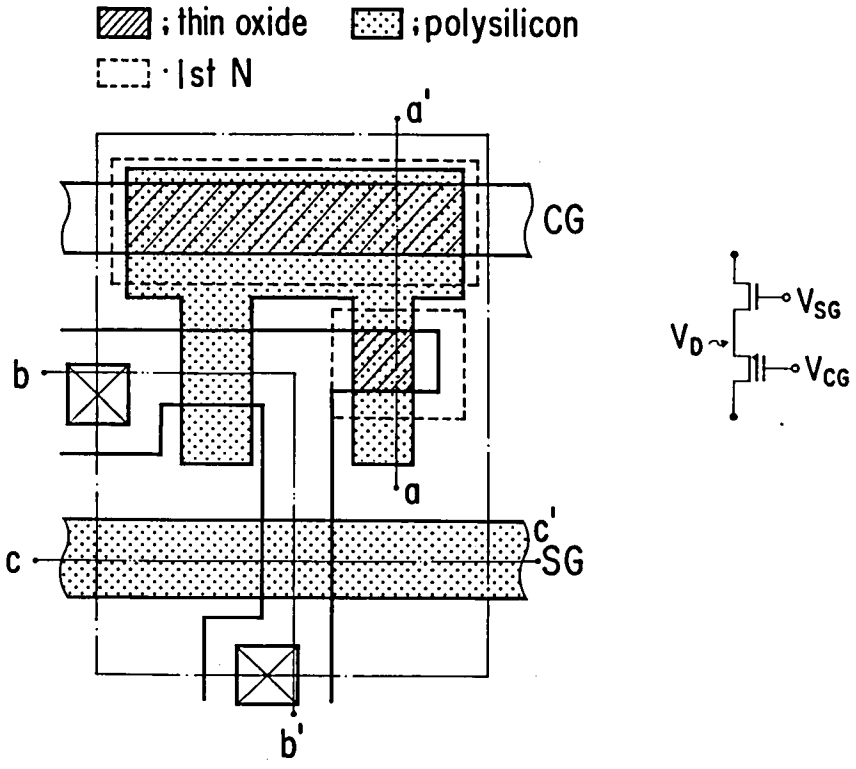


Fig. 4.11 Layout pattern of a new single-polysilicon EEPROM cell, named DIFLOX. The cell is composed of a selected gate and a floating gate transistors.

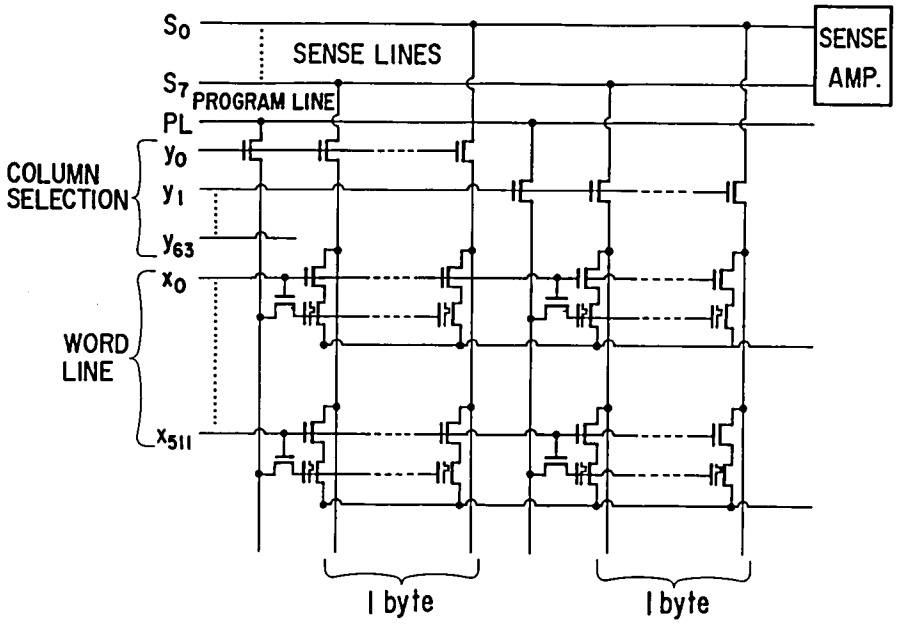


Fig. 4.12 Cell array of DIFLOX. The selected gate is equivalent to the word-line, and the control gate is connected to the program-line (PL) via a pass transistor.

applied to the cell, which can realize a field distance between the active regions as narrow as the minimum dimension determined by the photo-etching. In addition to this technology, a high voltage strucure was contrived to perform the erase/programming function. Figure 4.13(a) shows the cross section views of DIFLOX, along the lines of a-a', b-b', and c-c' in Fig. 4.11. As shown in Fig. 4.13(a), the high concentration N^+ below the thin oxide was covered with a lower doped zone of 0.2um thick. This structure improves the breakdown voltage of the PN junction. In order to optimize the distance between the N^+ control gate and the drain, the field inversion effect under the floating gate should be counted. By defining the turn-on voltage as the gate voltage at which the leakage current exceeds 10^{-10}A/um under a drain bias of 20V, the field turn-on voltages as a function of the length are measured as shown in Fig. 4.14. It is indicated that a 2um field length can guarantee the 15V erase/program operation. To eliminate the leakage current from the drain regions of the select-gate and floating-gate transistors to the substrate due to the avalanche breakdown, the source and drain of the select-gate transistors were doped with a low impurity concentration of $10^{17}cm^{-3}$, as shown in Fig. 4.13(b). The value is about one order of magnitude lower than the usual lightly doped drain(LDD), used for 1.2um NMOS FET for 5V circuits [43]. For the contact resistance reduction between the aluminum and the N^- region, a high dose implantation of arsenic was carried out through the contact

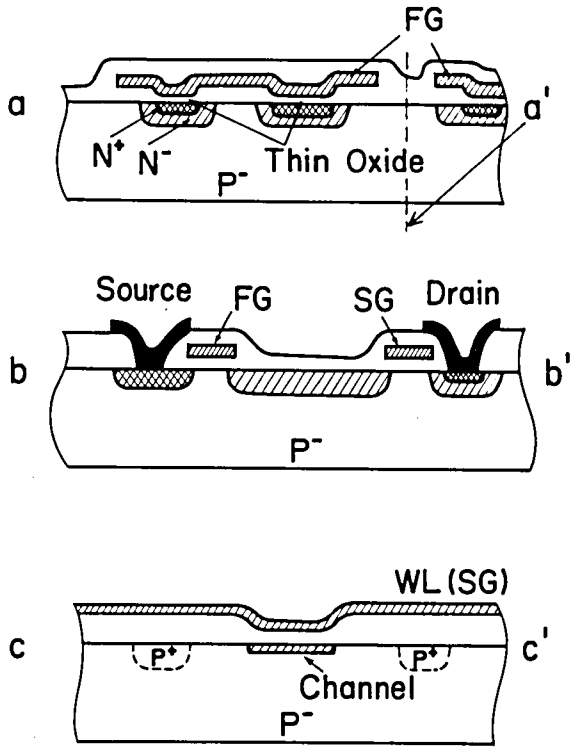


Fig. 4.13 Cross section views of DIFLOX, along a-a', b-b', and c-c', in Fig. 4.11.

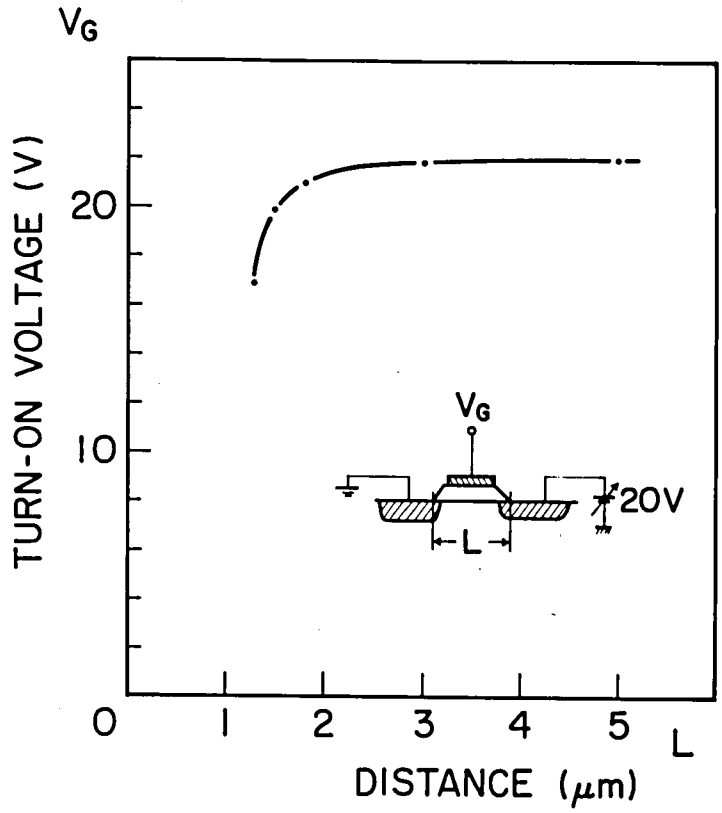


Fig. 4.14 Turn-on voltage of the field transistors as a function of the distance. It is defined as the leakage current exceeding 10^{-10} A/um. The drain is biased to 20V.

holes. This self-aligned formation of the diffusion area contributes to the cell area reduction. Only the source of the floating gate transistor was formed by N^+, as is used for 5V circuits.

For the design of the peripheral circuits, the overall characteristics of the cell array of Fig. 4.12 should be analyzed. The V_{th} shift of DIFLOX cell in the erased and programmed state as a function of the selected-gate voltage is shown in Fig. 4.15. The thin oxide thickness, d_1, is 85Å, and a pulse of 15V height and 2ms width is provided to the program line or bit line. The Figure indicates that the word-line should be elevated over 20V to eliminate the select gate transistor effect on the wide threshold window. This high voltage can be achieved in the strucure, shown in Fig. 4.14, as it is. However, it is safer to take some margin, which has already taken for the control gate design. So, heavy channel-stop ion implantation was carried out between the select gate, as shown in Fig. 4.13(c).

The read current measured under the conditions that the program line, bit line and word line are 2.5, 2.5, and 5V, respectively, exceeds 100uA in the conductive state by employing 1.4um gate length, in spite of the series connection of two high voltage transistors. The endurance characteristics of DIFLOX are shown in Fig. 4.16. Although it uses larger thin oxide area than the conventional double polysilicon cells, no characteristics' degradation could be observed up to 10^5 erase/program cycle.

The charge retention characteristics would be

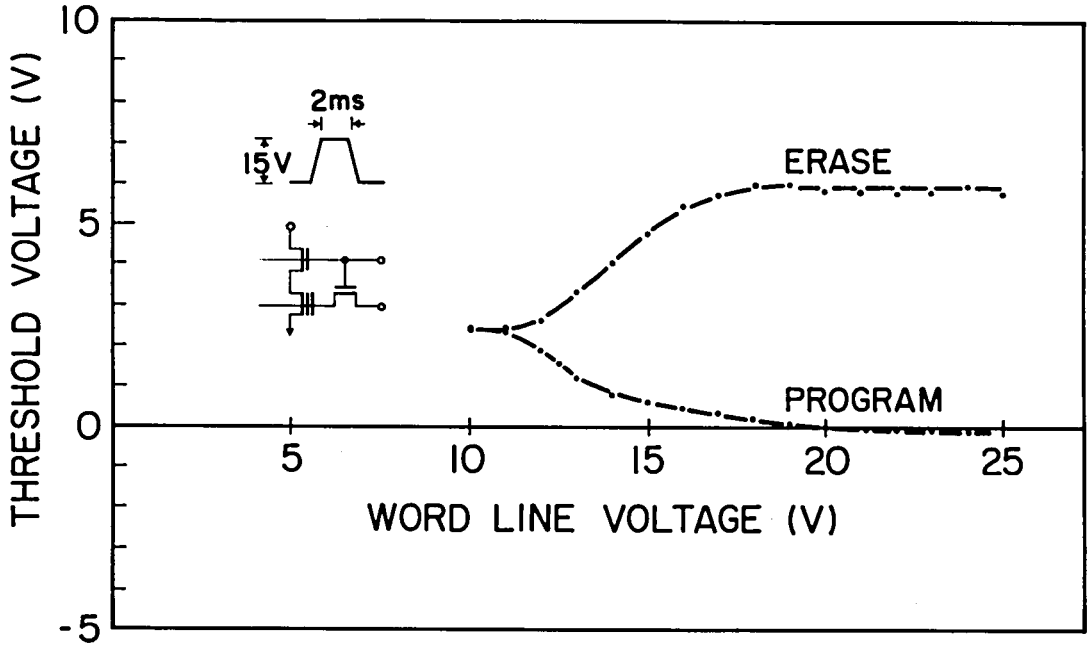


Fig. 4.15 Threshold shift of DIFLOX in the erase/program state as a function of a word-line voltage. V_{PP} = 15V, d_1 = 85Å.

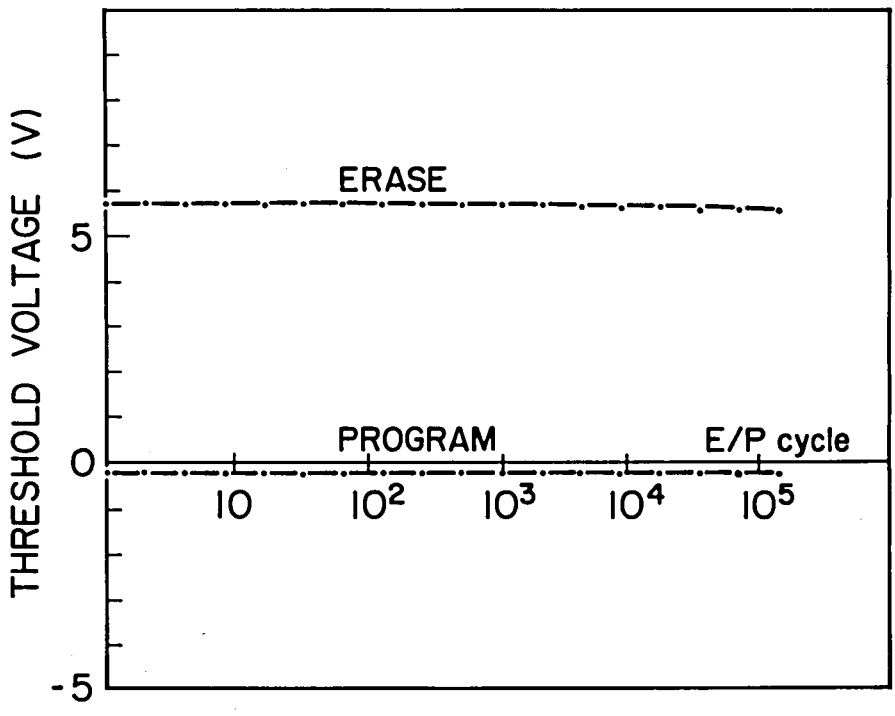


Fig. 4.16 Endurance characteristics of DIFLOX. V_{PP} = 15V, d_1 = 85Å, V_{WL} = 20V.

drastically degraded, if a thin oxide thickness of below 60Å were used, because of the direct tunneling effect of oxide. However, up to the thickness, the retention characteristics are mainly dominated by the barrier height of polysilicon oxide, surrounding the floating gate, rather than that of thin oxide[11]. So, the cell can guarantee the same charge retentivity as the conventional UV EPROM cell.

The microphotograph of the DIFLOX cell array is shown in Fig. 4.17. The node labels correspond with those of the equivalent circuit illustrated below. WL, PL, and SS are the word line, program line, and cell source, respectively, which have already appeared in Fig. 4.12. B_i (i=0,1,...,7) means a bit-line connected to the sense line S_i via the column selection transistor, as shown in Fig. 4.12.

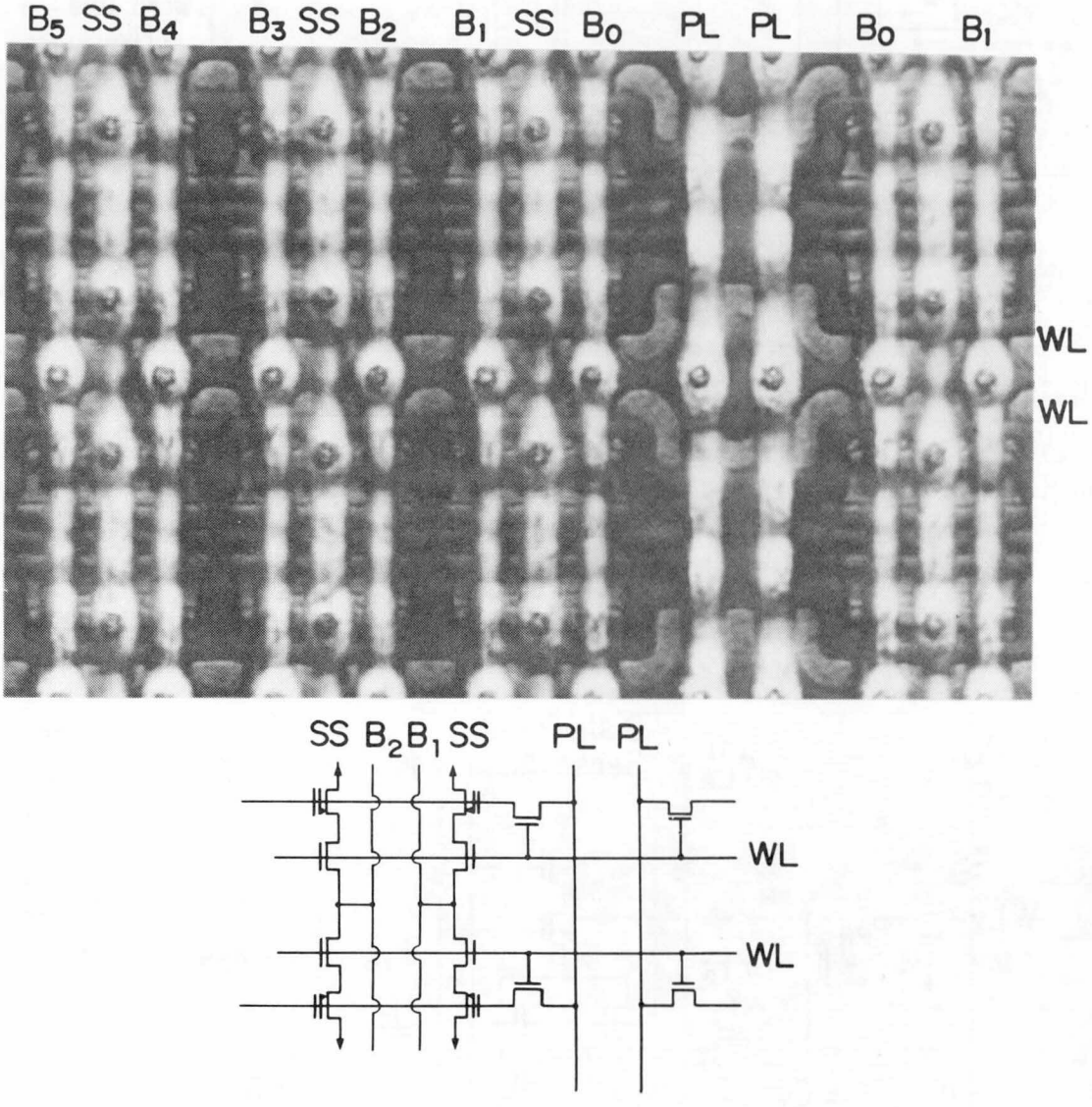


Fig. 4.17 Microphotograph of DIFLOX cell array. The cell size is 7.5*11.5um.

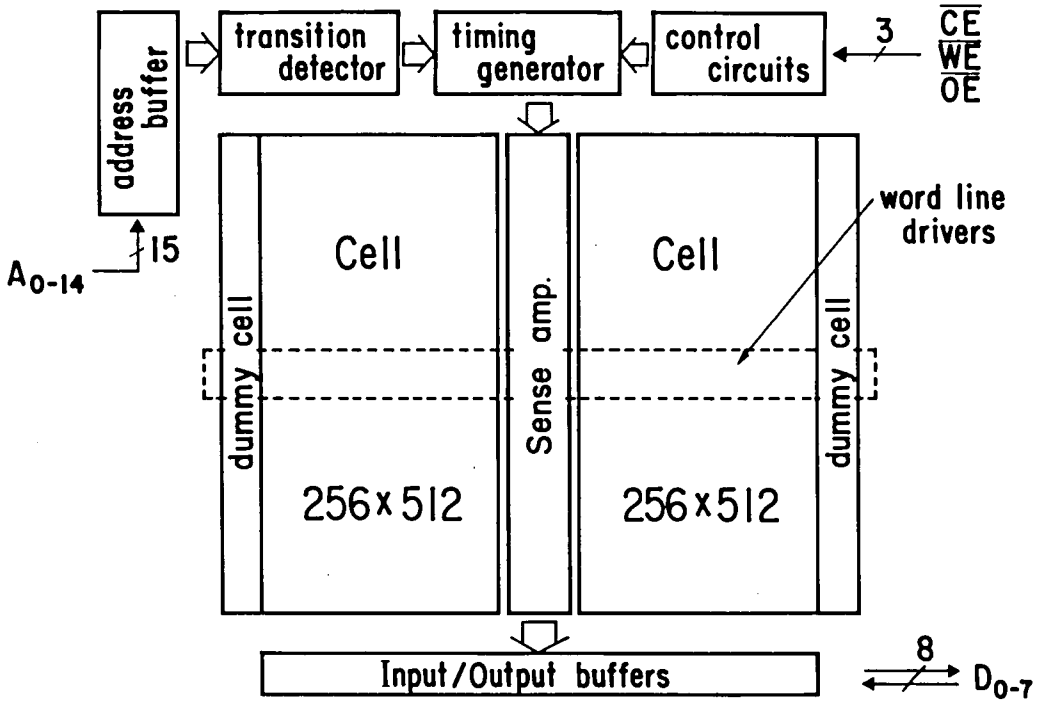


Fig. 4.18 Block diagram of 256kbit EEPROM in the read operation.

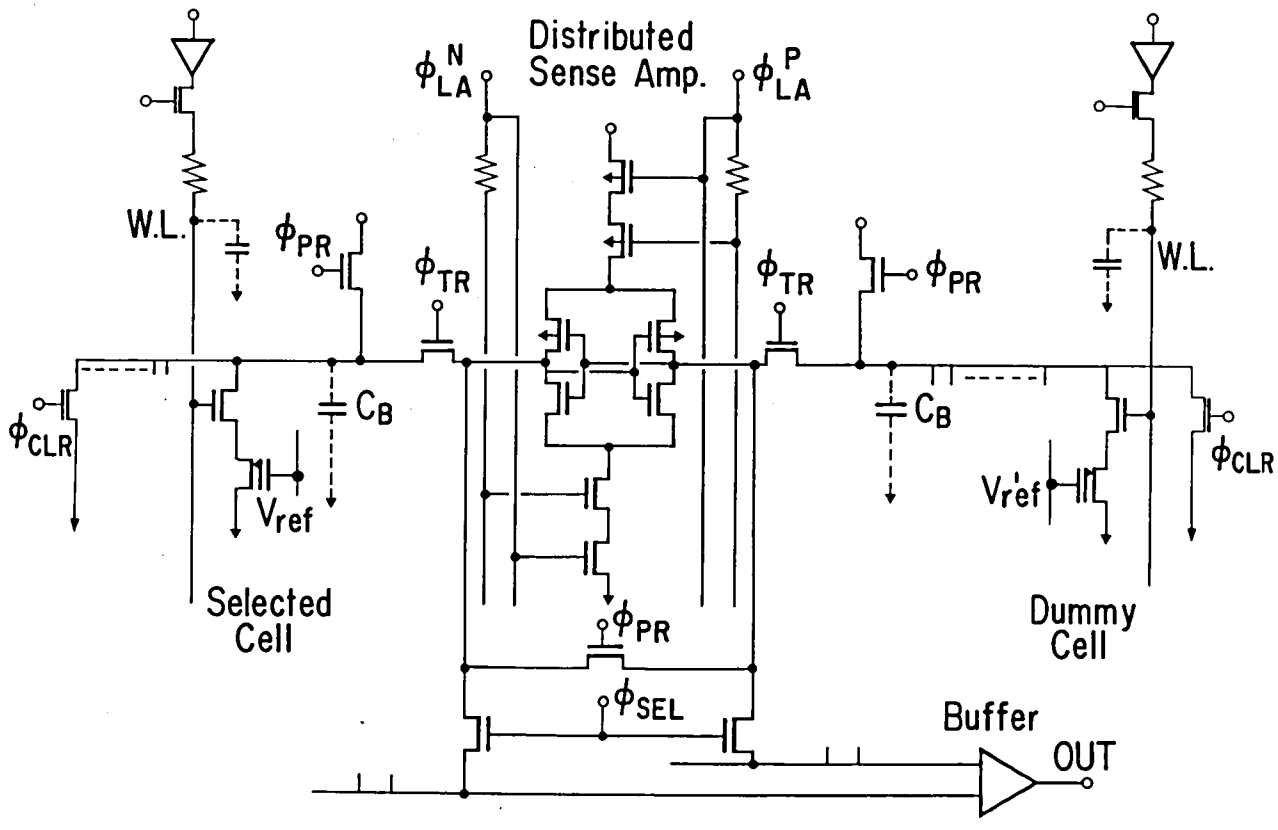


Fig. 4.19 Cell peripheral circuitry. Improved open-bit-line scheme and distributed sense amplifiers are shown.

4.4 Open Bit-line Structure

In order to compensate the large bit-line parasitic capacitance, an open bit-line scheme which has been already used for current DRAM design, has been improved and applied to the EEPROM. The block diagram of the memory in read operation is shown in Fig. 4.18. The cell array is divided into left and right hemispheres, and the sense amplifiers are located between them. Every bit line has 256 cells and one dummy cell. When the memory cell on a hemisphere is selected, the dummy cell on the other side is selected as a reference. The cell peripheral circuitry in the read operation is shown in Fig. 4.19. At the transition edge of the address input signals, the precharge pulse, Φ_{PR} is generated internally. It precharges bit-line pairs and their corresponding sense-amplifier nodes, and equalizes them simultaneously. The successive clocks triggered by Φ_{PR}, control all of the reading procedure as follows. In the period of the prechrge, the bit lines are raised to V_{DD} - V_{th}. The level is estimated to be 2.5V by the body effect of NMOS FET. All word lines including the dummy cells' are biased to zero. When Φ_{PR} goes low, the selected-cell and the dummy-cell word lines are raised simultaneously. So, both bit-line levels begin to fall gradually, according to the conductance of the selected and dummy cells. As no pull-up devices are activated, the level variation in this free-running state is expected to be fast. In addition, the bit-line parasitic capacitance equally contributes to the fall time of both bit-lines. After an appropriate period,

Φ^{P}_{LA} goes low, and then Φ^{D}_{LA} goes high. In other words, the sense amplifiers are activated with PMOS flip-flops first and NMOS second. The bit line data are boosted by them. Finally, Φ_{IR} goes low, and the bit-lines are isolated from the sense amplifiers. In the quiescent state Φ_{DR} is high. So, bit-lines are biased to zero in order to eliminate the idle current and to avoid cell data destruction by noise. As the voltage of the program line (PL) shown in Fig. 4.12, keeping about 2.5V during the read operation, transfers through two pass transistors controlled by the column selection, y_i and the word line selection, x_j, 2.5V bias is applied as a V_{rd} shown in Fig. 4.19. The bias is provided to the control gate of 8 selected cells. These internal timings of read mode were verified by the electron beam tester, as shown in Fig. 4. 20, although the voltage levels are not correct because of the measurement disturbance by the neighbor pattern's potential.

It is important for open-bit-line scheme to synchronize the bit-line signal variation associated with the word-line delay to the activation of sense amplifiers. The delay can be estimated by assuming 256kbit cell array to be divided into 4 blocks. The parasitic word-line capacitance including the gate and the substrate is 0.85pF, and the resistance between both ends is 41kΩ. The worst-case RC delay is calculated as 35ns, resulting in about 1.0V of bit-line changing. So, a scheme of distributed sense amplifiers was applied as shown in Fig. 4.19. The activation of sense amplifiers is designed to depend on the

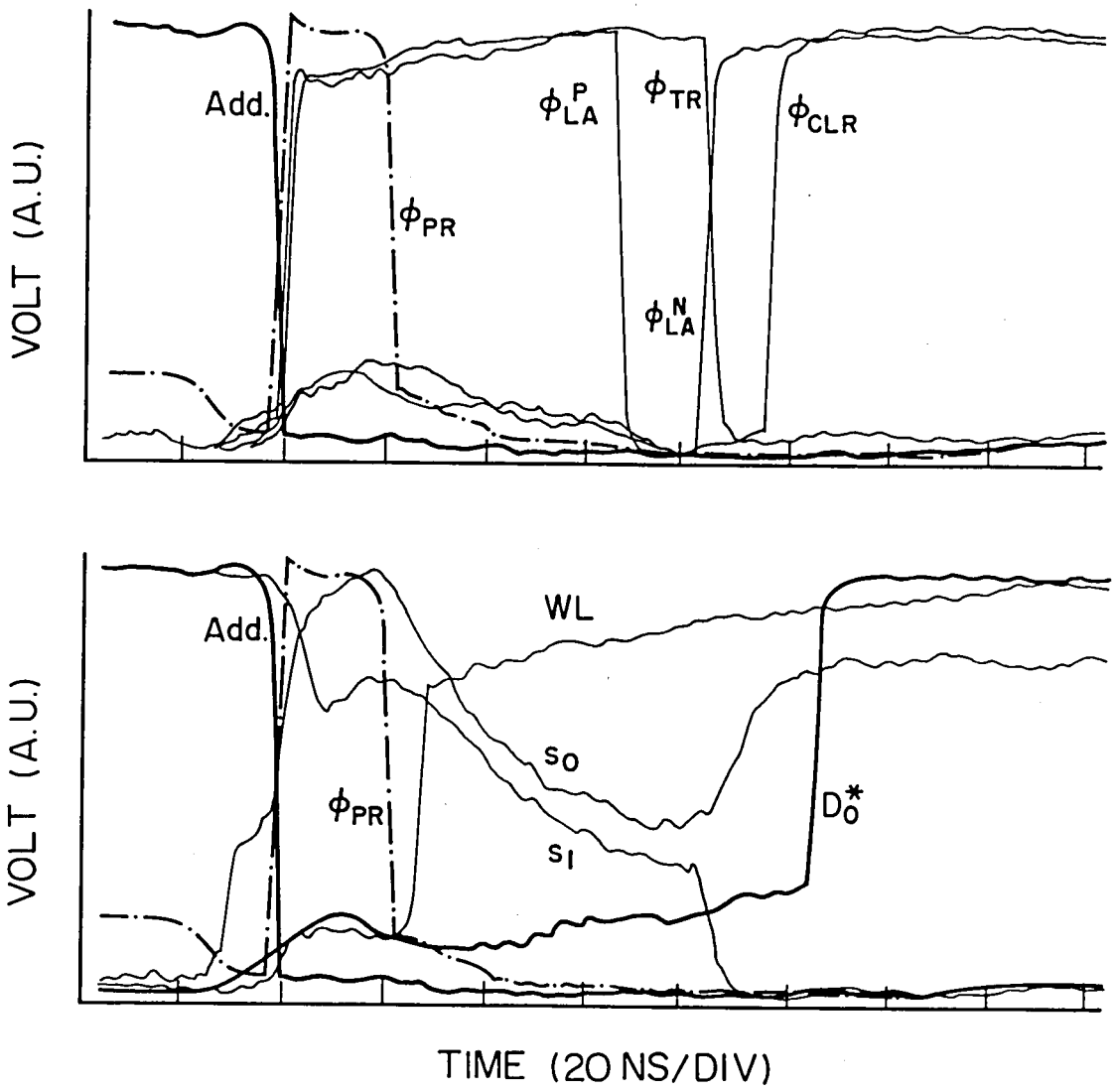


Fig. 4.20 Monitored waveforms of internal nodes, obtained by the electron beam tester. s_0 and s_1 are two nodes of a sense amplifier, and D_0^* corresponds to the output signal before the final buffer. The levels are incorrect, not only absolutely, but relatively, because they are influenced by the electron scattering of the neighbor patterns in the LSI.

distance from the word line drivers, by adjusting the gate capacitance and the line resistance. It is accomplished by making the word-line and the sense amplifier activation line equivalent or have same loads. This enables the bit line data to be latched just at the time when the voltage difference reaches an appropriate level, wherever it may be located. It also suppresses large peak current caused by the sense amplifier activation, and makes it broad.

It is necessary for the non-volatile memories that the logical data coincide with the physical data, which is different from the DRAM design. The necessity also owes to the attribution of EEPROM cell architecture, that is, only programming can be performed selectively, while DRAM cell can be written both "1" to "0" and "0" to "1", individually. Therefore, the final output data are inverted on the buffer, if it comes from the left hemisphere, while the data from the right hemisphere comes out as it is. So, the data in the chip clear state, which is defined all the cells in positive threshold shift, correspond to all one.

4.5 Erase/Program Control Circuits

The page-mode programming gives one solution to reduce the erase/program (E/P) time. This feature applied to the 256kbit EEPROM allows programming up to 16 bytes (128 bits) simultaneously. The scheme fits into the open-bit-line structure, since the flip-flop which acts as a bit-line sense amplifier in the read mode plays a role of a temporary storage for programming data, as shown in Fig. 4.21. Therefore, the data can be written in almost the same timing as a conventional SRAM. The timing chart of E/P procedure is shown in Fig. 4.22.

The erase/program function is explained as follows, using Figs. 4.21, and 22. The "pump" means the high voltage generator, or the charge pump, whose characteristics will be given in the next paragraph. Once the chip detects the write enable, $\overline{WE}$, going low, all the sense amplifiers are activated and isolated from the bit lines by $\overline{\Phi_{IR}}$. The programming data are loaded to the latches (sense amplifiers) successively, during the low $\overline{WE}$. If the time, 100us, lapses without any low $\overline{WE}$ inputs, the chip goes into the erase state automatically. In the erase, the bit-lines are biased to zero by Φ_{DR}. And both the program line (PL) and selected word-line are raised. As a result,the control gates of the selected eight cells are elevated to V_{PP}. Consequently, the electrons are injected into the floating gate, and selected cells get into the erased state. Next, the chip turns into the program. Φ_{IR} goes high and the stored data are transferred to the bit lines. The charge

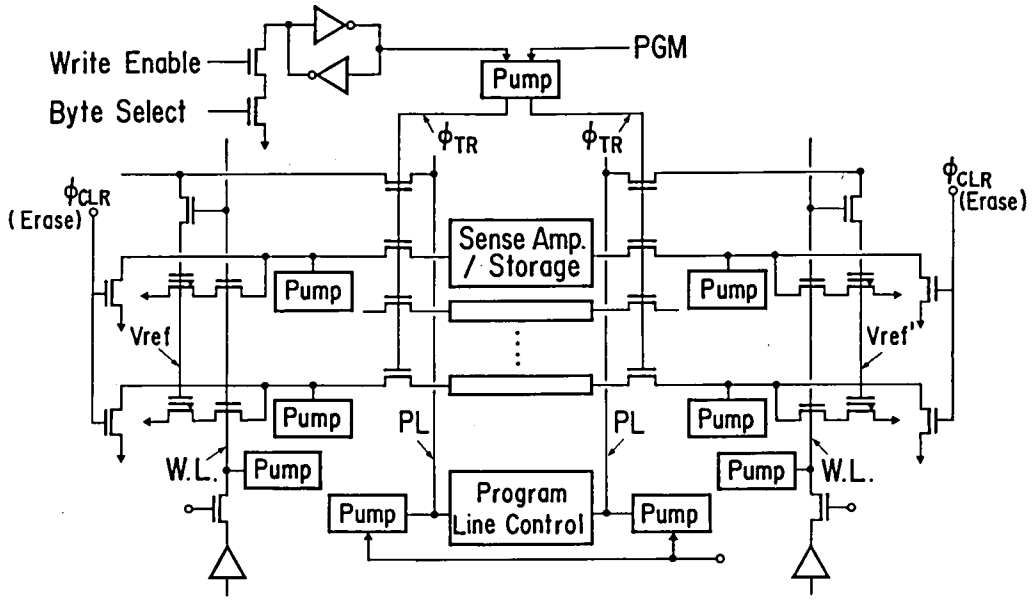


Fig. 4.21 Cell peripheral ciruitry in erase/program operation.

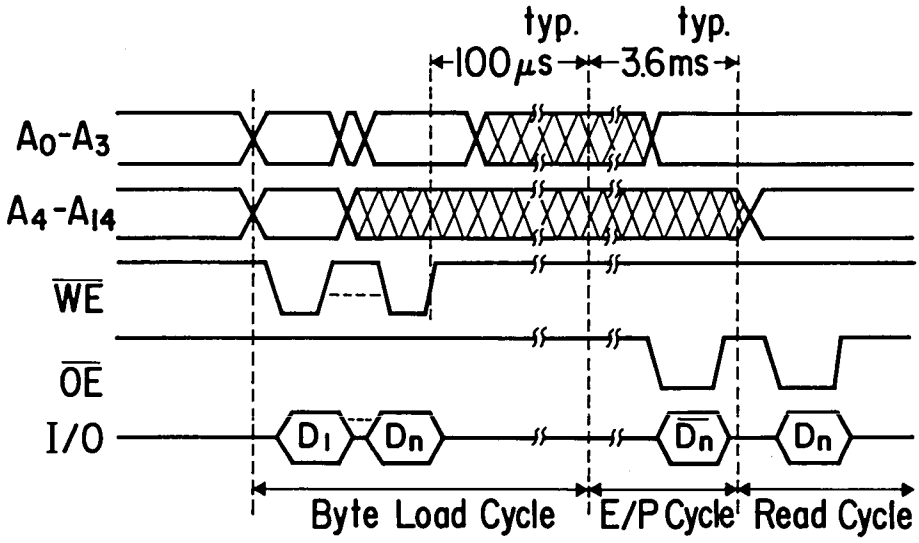


Fig. 4.22 Timing chart of the page-mode programming and the data polling.

pumps elevate bit-lines of logical high (5V), to V_{PP}, while the bit-lines of logical low (0V), keep 0V level, because the drivability of charge pumps is sufficiently smaller than that of the pull-down latch transistors. As the PL is biased to zero, the programming is performed on the corresponding cells. Only in the drain voltage raised cells, the stored floating gate charges are released to the drain. As each byte storage has one additional flag latch which is set when the data is loaded, the E/P procedure is performed only on bytes that are actually being programmed.

The data polling function, also shown in Fig. 4.22, is accomplished by transferring the stored data to the output through sense lines during the E/P state, if the output enable, $\overline{OE}$, is low. As a result, the data are read inverted from the data to be stored. After the programming finishes, the reading procedure automatically starts, and right data are sensed and stored.

The timing circuits constructed by the switched-capacitor technique take care of the complicated E/P procedure. This technique is widely applied to recent MOS analog circuits such as digital filters [4,40,41]. The lapse of the 100us writing time for successive $\overline{WE}$ pulses is controlled by the Φ_{START} generator circuit as depicted in Fig. 4.23. At first, the V_{DD} should be over 4V for any erase and program occurrance. The function is necessary to avoid the unintentional E/P by unstable voltage supply, or by misoperation. If the condition is satisfied, P_{ON} goes high. Next, the $\overline{WE}$ (Write Enable) comes. The rising and falling

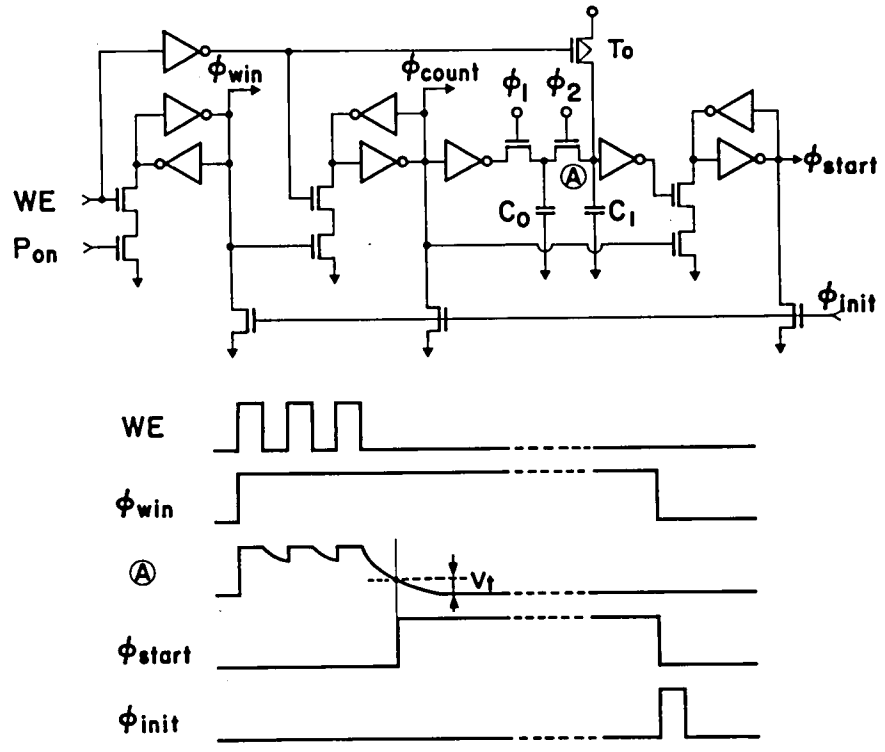


Fig. 4.23 Successive data loading control circuits (Φ_{START} generator). A period of 100us is counted by the switched capacitor.

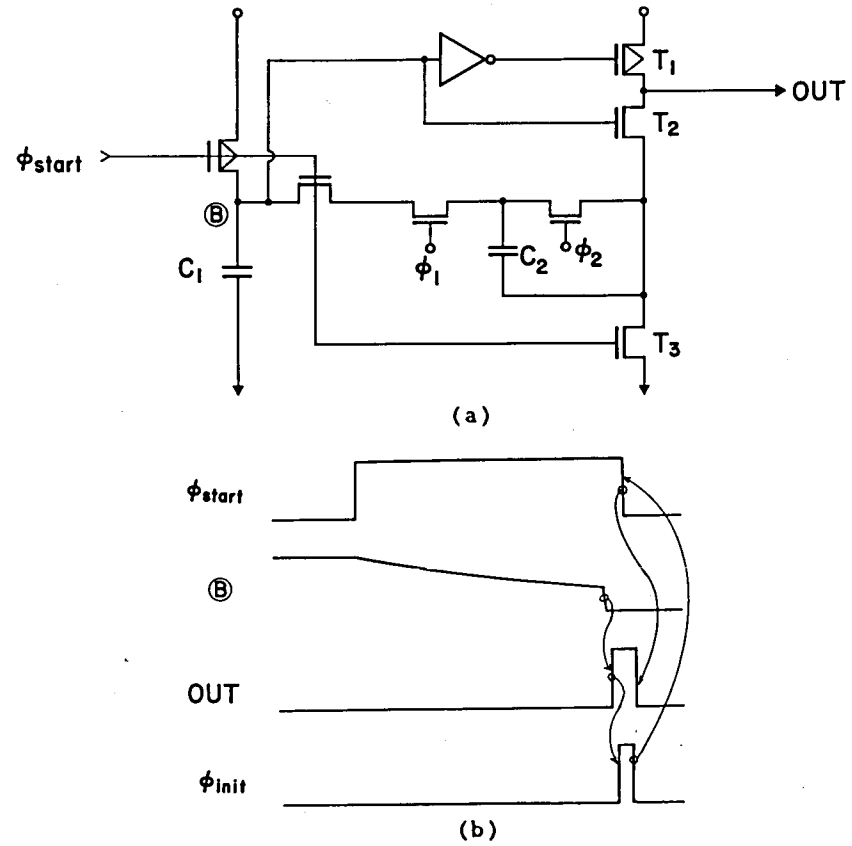


Fig. 4.24 (a) Timer circuit, and (b) wave forms at several nodes for erase/program. Improved switched capacitor is used. It counts 1.8ms lapse time.

edges of the first WE pulse generates two control signal generation of "Write procedure INitiate", pulse, Φ_{WRI}, and "start COUNTer" pulse, Φ_{COUNT}. The pulse, Φ_{WRI}, initiates the status flip-flops such as flag latches incorporated with each byte. The clock, Φ_{COUNT}, activates internal oscillator of 10 MHz and raises V_{PP} by charge pumps. This basic clock goes to some high voltage pumps, and makes them ready. It also generates non-overlap pulses, Φ_1 and Φ_2. The operation of the Φ_{START} generator is explained as follows. Every time the $\overline{WE}$ pulse becomes high, the capacitor C_1 is charged by the transistor T_0, and the node Ⓐ in Fig. 4.23 goes to high. If "$\overline{WE}$" keeps low, the node Ⓐ falls gradually due to the multiple dischargings by the smaller capacitance C_0. The RC time constant of this falling is calculated from the switched capacitor theory as C_1/fC_0, where f is the frequency of Φ_1 and Φ_2. At last, the node Ⓐ reaches the turning voltage, V_t, of the next inverter, At that time, Φ_{START} is generated. This Φ_{START} makes all E/P circuits active and initiates the second timer which controls much longer time for E/P. The circuit is shown in Fig. 4.24(a), and the wave forms of nodes are illustrated in Fig. 4.24(b). The second timer designed by a modified switched-capacitor technique counts milli-second order erase and program time using within the reasonable capacitance occupied silicon area. In contrast with the conventional one, the voltage drop through the discharge capacitor C_2 is determined by the conductance ratio of transistors, T_2 and T_3, the capacitance area was reduced to about one tenth in

comparison with the conventional technique. At the end of E/P, those timers generates Φ_{INIT} which resets all the flip-flops in the Φ_{START} genetator as shown in Fig. 4.23.

Figures 4.25 and 4.26 show the drivability characteristics of the internal high voltage generator and the pumping circuits. To provide sufficient current, the high voltage generator shown in Fig. 4.25 has a comparatively low output impedance. The output terminal is clamped by the series connection of diodes for regulation. The rise time of this high voltage does not affect the thin oxide stress. On the other hand, the pumps shown in Fig. 4.26, which are attached to every bit and word line, are designed to have relatively smaller drivability in comparison with the parasitic capacitance in bit- and word-lines, which makes the voltage rising speed slower. The rise time was measured as 50us, which is enough to reduce the thin oxide stress.

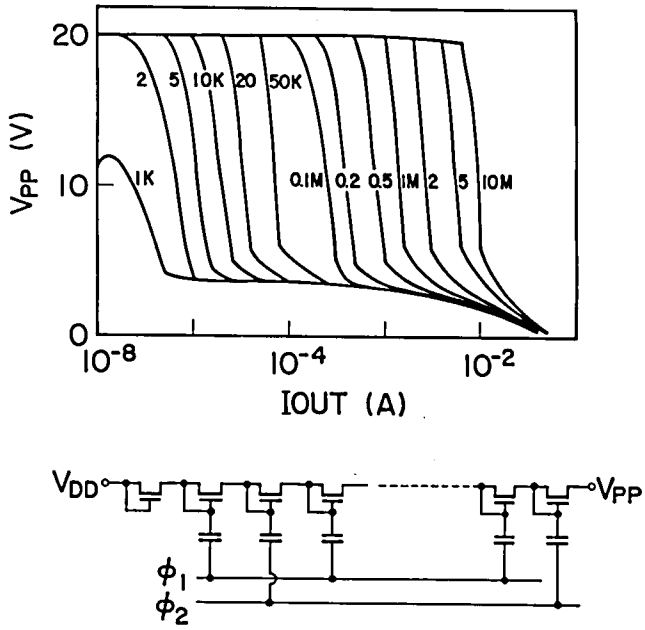


Fig. 4.25 Output characteritics of high-voltage generator. In the memory, the basic frequency by the internal oscillaor is set to 10MHz.

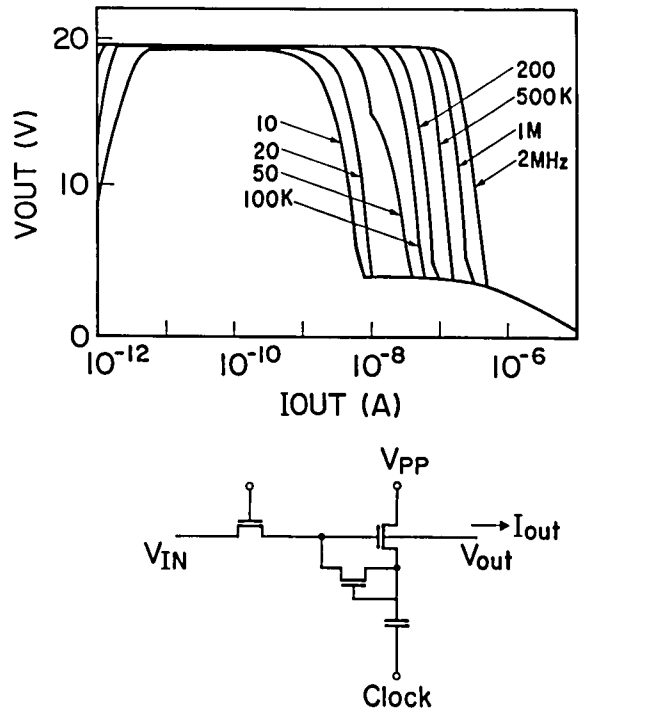


Fig. 4.26 Output characteristics of high-voltage pump, It is attached to every bit and word line.

4.6 Performance and Results

The microphotograph of the 5V only 256kbit CMOS EEPROM is shown in Fig. 4.27. The column decoder and sense amplifiers/latches are vertically placed at the center of the cell array, while the word line decoder/drivers are located horizontally. The large capacitance array, the component of the high voltage generator is shown at the right bottom side. The die size is 6.23*7.33mm and is packed into the DIP 28pin standard package. It is pin-to-pin compatible with the current 256kbit SRAM [43]. The chip contains several circuits which enable the screening of the chip. One is "clear", which makes all the cell erase state at one time. The second is the read margin check which enables the gate voltage of the cells variable. These are activated by providing 12V to appropriate pins.

The monitored waveforms of the internally generated V_{PP}, the erase/program control signal, Φ_{EP}, and the input write enable, $\overline{WE}$ are depicted in Fig. 4.28. The dislocation on the center of Φ_{EP} shows the changing from the erase to program state. It indicates that V_{PP} is elevated to 20 V, and all the timings are well-controlled corresponding to the page-mode programming. The total amount of the erase and program time is 3.6ms/16byte at 5V power supply.

The photograph of the address and data out waveforms with external 100pF load is shown in Fig. 4.29. The address access time of 150ns was successfully obtained with a normal 5V voltage supply. The active power dissipation is 80mW. Similar to the other CMOS devices the active power

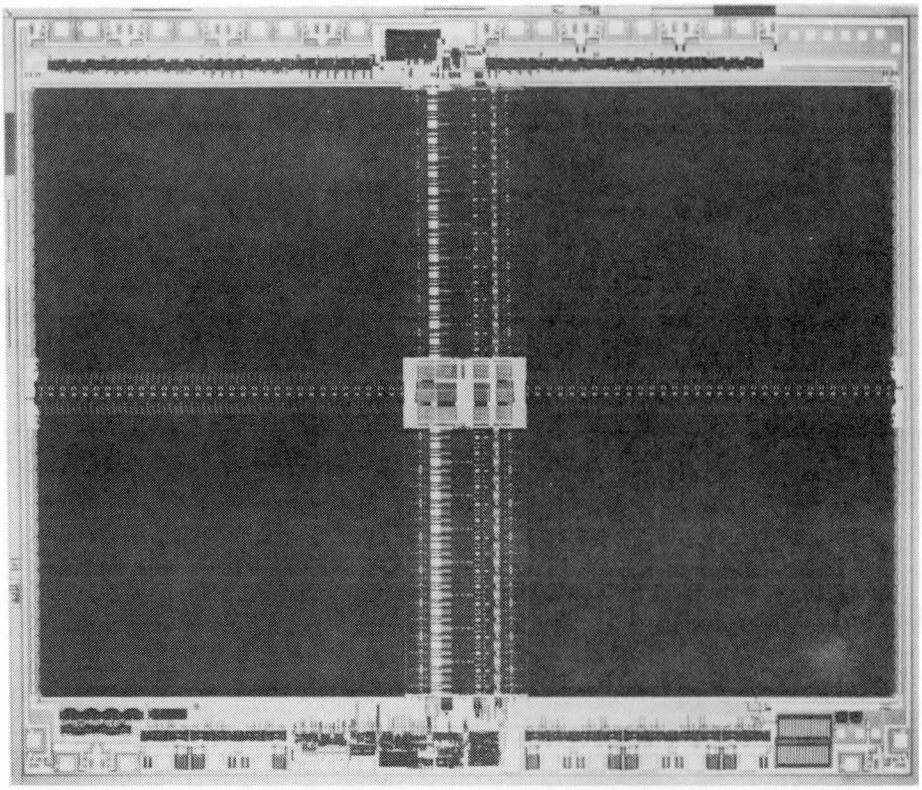


Fig. 4.27 Microphotograph of the 256kbit EEPROM whole die.
The die size is 7.33*6.23mm.

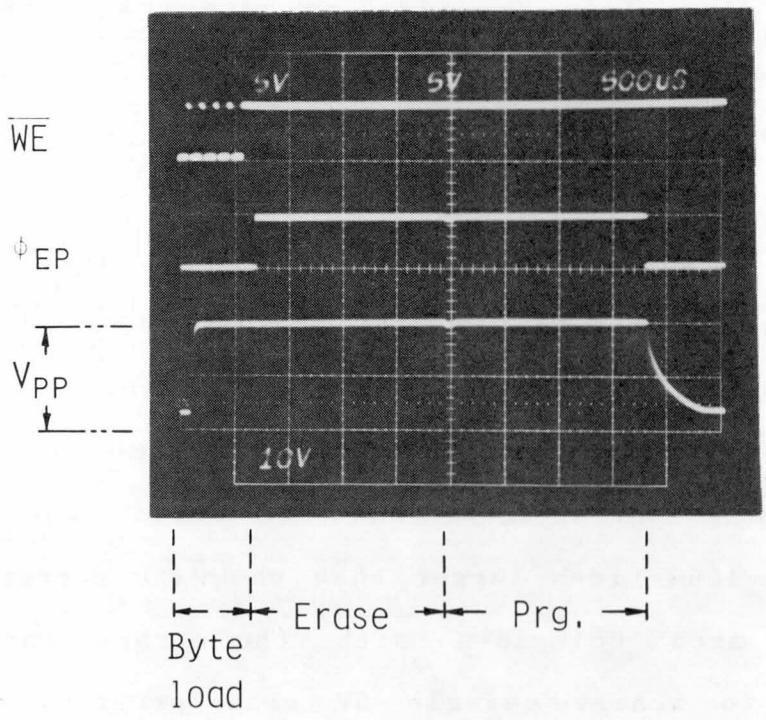


Fig. 4.28 Monitored waveforms of Φ_{SA} and V_{PP}.
V_{PP} is raised to 20V.

dissipation depends on the read cycle time. It reduces to 35mW at a 1MHz read cycle, and drops below 1uW at the standby mode.

The characteristics of this EEPROM are summarized in Table 4.2.

4.7 Summary

The single polysilicon EEPROM cell was useful for the logic combined LSI. Now, it is verified to be quite feasible for the high density standard EEPROM, if the stacked area of the diffused control gate and the floating gate is reduced by using the thin oxide as a dielectric. The charge loss from the floating gate toward the control gate through the thin oxide does not affect to the erase/programming characteristics, as long as the threshold voltage shift is within 5V. Two single polysilicon cells have been proposed. However, as long as the 1.2um CMOS technology is applied, the drain thin oxide area should be separated from the floating gate transistor, although the cell area is larger. Because, under the same electric field between the floating gate and the drain, the substrate current is 1000 times larger than the gate current, if the tunneling area coincides with the transistor. It is difficult to achieve single 5V erase/programming by the

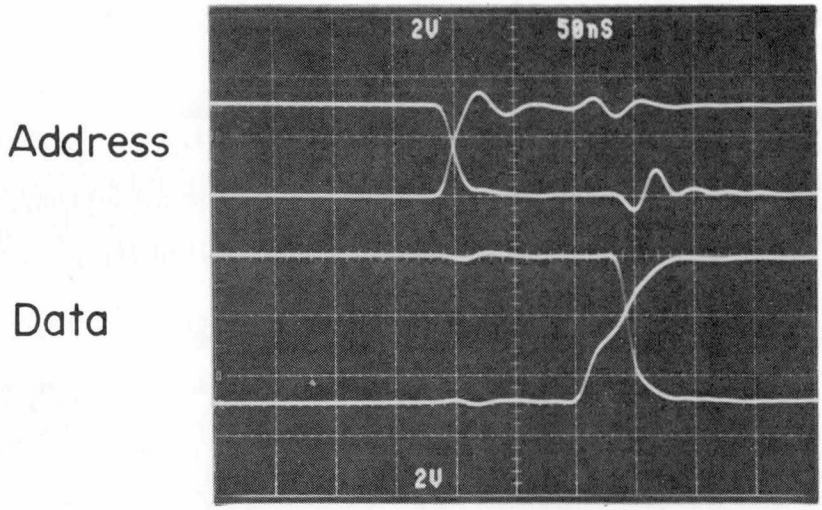


Fig. 4.29 Oscillograph of the address and data-output waveforms with external loads.

Tab. 4.2 Characteristics of 5V only 256kbit EEPROM.

Organization	32kword*8bit
Operation	Fully asynchronous
	(with internal clocks)
Technology	N-well CMOS
	(single poly & single Al)
Minimum gate length	1.4um
Gate Oxide thickness	250Å
Thin Oxide thickness	85Å
Cell size	86.25um
Die size	7.33*6.23mm
I/O interface	TTL
Address Access Time	150ns(typical)
Chip Select Access time	150ns(typical)
Erase/Program Time	3.6ms(typical)
	(16byte page)
Active Power	80mW
	35mW(1MHz)
Standby Power	1uW
Package	28pin DIP
	(256kbit SRAM compatible)

later one. However, the choice quite depends on the available technology.

A 5V only 256kbit CMOS EEPROM employing a new single polysilicon cell, DIFLOX, has been successfully designed and fabricated. Exploiting 1.2um design rules, and 1.4um gate length for the high voltage transistors, the cell size was reduced to 86.25um^2. The cell was designed to eliminate any leakage current caused by the junction breakdown and the field inversion. The cell, which offers a 5V threshold window between the erase and program state, allows 100uA reading current in the conductive state. It guarantees over 10^5 erase/program cycles. By using the page-mode programming, the EEPROM allows successive data load up to 16 bytes with the same timing as an SRAM. The erase/program is carried out for the bytes at once within 3.6ms. All the timings are controlled by several kinds of clocks, all of which are generated internally by the timers constructed by the switched capacitor technique. The improved open bit-line scheme was applied for the read operation. By using the distributed sense amplifier scheme for the word-line delay compensation, 150ns typical address access time was obtained under 80mW power dissipation. This internally synchronous and externally asynchronous architecture fits the page-mode programming, because the sense amplifiers are also used to share the internal storage of the programming data. Due to the CMOS circuit configuration, the chip offers extremely small standby power dissipation.

Chapter 5

Design of an Application Specific Memory IC for LSI Function Testing

5.1 Overview

Recently, we have seen an explosion in the development of application specific ICs, (ASICs) for use in systems ranging from computers to communications. This increase in integrated circuit designs has been driven by the improvements in both IC technology and computer tools. However, testing these chips after they are manufactured remains a serious problem. As more complicated functions have been integrated in one VLSI chip, the cost of testing has become a significant part of the total cost. To reduce testing costs, designers often include circuits to improve testability of the chip[44,45]. Some people have included enough hardware on chip so they are able to performance speed testing using a low speed test setup [46,47]. However, performance testing usually depends on expensive testers, which contain a sophisticated controller, high speed vector RAM, large test heads with high speed pin electronics, and large wide cables to connect everything up. These complex testers may be necessary for parametric testing and for final production testing but they are not well suited for initial engineering debug and evaluation. For these applications a more flexible setup is needed, preferably one that is inexpensive enough that it can be allocated to a single project.

This chapter describes the Data Generator and Receiver (DGR), an architecture for a chip that can be used to build a small, flexible, low-cost functional tester. The chip contains 6K - 8Kbit vector memory besides the sequencer and pin electronics. The one-chip memory gains data traffic band width, enabling high-speed testing. So, DGR is a kind of application specific memory ICs (ASMIC).

Using integrated circuit technology to help solve the tester problem has many advantages. The resulting tester is physically small allowing it to be integrated directly on a probe card or a test board. All the high speed signals would then be confined to very short wires between the Device Under Test(DUT) pins and the tester chips; the long cables between the tester and the host computer could operate at much slower speeds. The performance of the tester chips should scale with the device performance, since they both can be manufactured with same technology. Finally since the entire tester consists of a small number of identical chips, the overall cost of the tester should be low. This lowcost would make it possible for individual groups to have their own tester, rather than having to time-share on a large tester.

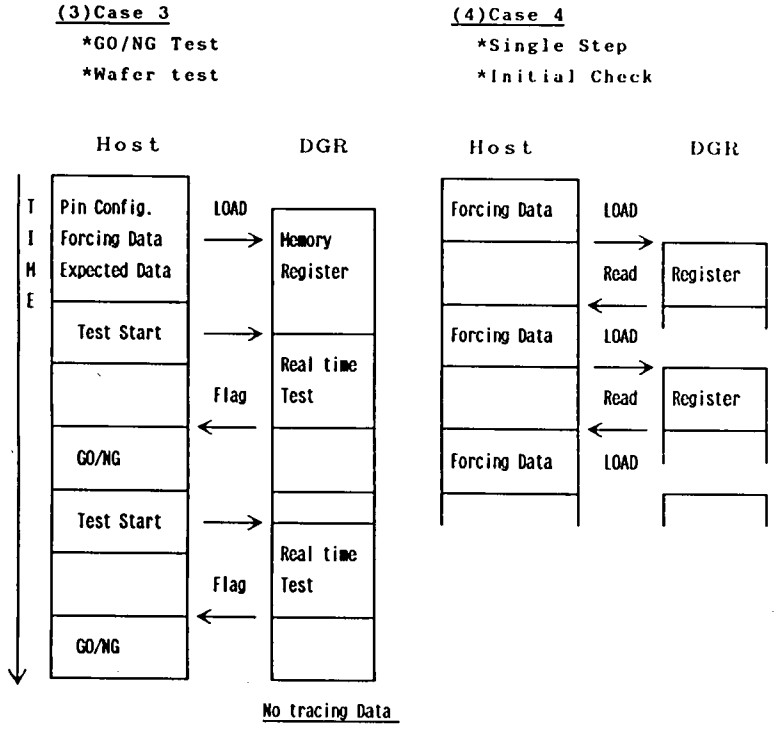
Section 5.2 and 5.3 describe DGR system design and on-chip memory design, respectively. In Section 5.4, some circuit techniques implemented to DGR are explained. Section 5.5 addresses DGR design environment. Section 5.6 describes the performance of the prototype DGR. Some results obtained by the refined one, fabricated by the advanced technology are given. The work is summarized in Section 5.7.

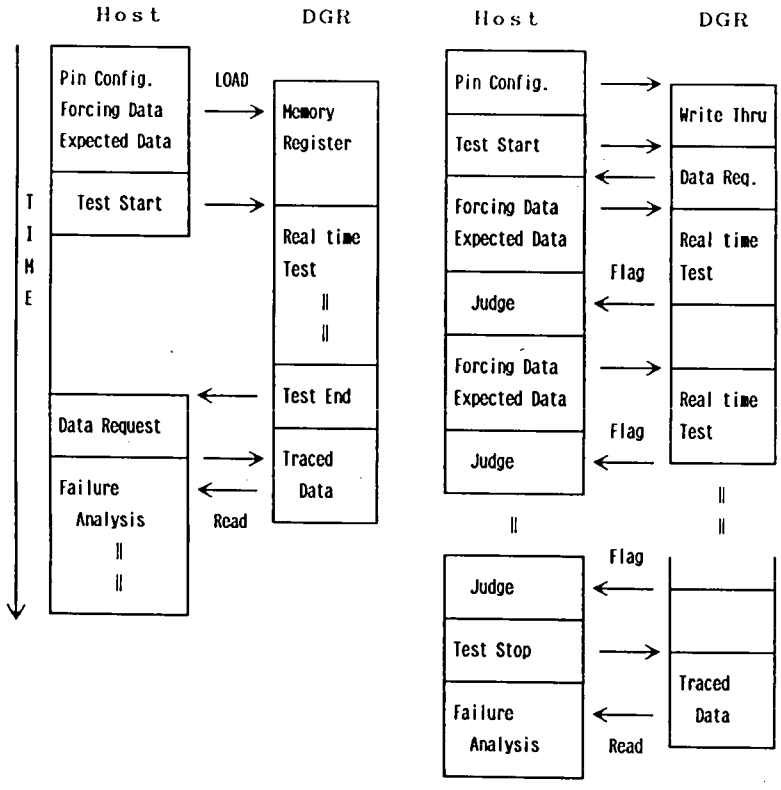
5.2 DGR System Design

The DGR was designed to be a basic building block for a functional tester. Each chip would drive a number of DUT pins and contain memory to store the test vectors for these pins. For the prototypes that was designed, each chip drives 16DUT pins and stores 256 vectors. The number of tester pins can be expanded by running multiple DGR chips in parallel, and the vector depth can be expanded by cascading DGR chips. Figure 5.1 shows an example tester built using DGR chips. During the setup phase of the test the host processor loads the test vectors, sequence information, and personality of the DUT pins. This interface is completely asynchronous; the DGR chips look like conventional SRAM to the host. After the test has been loaded the "BeginTest" pin is asserted and the DGR runs the test using the stored vectors to both drive pins and check the results. The chip contains a set of registers to store the addresses of vectors where errors are detected. During a test the DGR can also use its vector memory to store the actual values on the DUT pins. The contents of the memory can be read after the test to help the designer debug the device being tested.

Each test vector contains two bits per pin. One bit, the "state bit", determines whether the tester should be driving or sensing the DUT pin. The other bit, "the value bit", sets whether the pin should be high or low. The value bit sets the output level when the pin is a driver, and is used as the expected value when the pin is a receiver. The

Tab. 5.1 Test sequences.





use of two bits per pin means that the DGR chip can easily handle bidirectional I/O pins. Since all pins can be bidirectional, the user need not worry about connecting input pins of the DUT to the "right" DGR pins. This means a standard tester board can be used to test a number of different chips. The only custom wires that need to be provided by the user are power supply connections.

The DGR also supports a lower speed read/write through mode for tests that require a very large number of vectors. In "Write-Through Mode" the test vectors are provided one at time through the host interface, synchronized with the DGR by using the "DataRequest" pin. Since the vectors are supplied externally, there is no limit on vector length. The maximum vector rate in Write-Through Mode will most likely be limited by the connection between the host and the tester, and thus there is a trade off between vector length and maximum vector rate.

The several kinds of test sequences using the DGR are summarized in Tab. 5.1.

5.2.1 Block Diagram

As is shown in Fig. 5.2, the DGR can be divided into three main blocks: address control, vector memory, and DUT pin control, The heart of the chip is the vector memory which stores the patterns to apply the DUT pins. The memory has two different modes of operation. In the normal operating mode the value bit storage of the vector memory is used twice. At the beginning of the test it stores the expected value of the DUT outputs. This value is compared with the actual DUT output and generates an error if they do not match. During the test, these values are overwritten with the actual DUT pin values, allowing the host to read back the actual values after the test. For wafer probe this overwriting of data could increase the testing time , since it would force the host to reload the vectors after every test. To avoid this problem, the DGR provides a memory protect mode. In this mode the vector memory can only be changed by the host; running the test will not cause any of the value bit change. Even in the write protect mode, all the informations about the error, the error addresses, and error bits, are kept in the registers, as explained in the following paragraphs. So, the host can easily recognize how was the DUT.

5.2.2. Pin Drive

The output of the vector memory is sent to the DUT pin drive block. This circuit controls the format and timing of the outputs and inputs. For each DUT pin, there are two eight-bit registers Areg. and Breg., that are used to

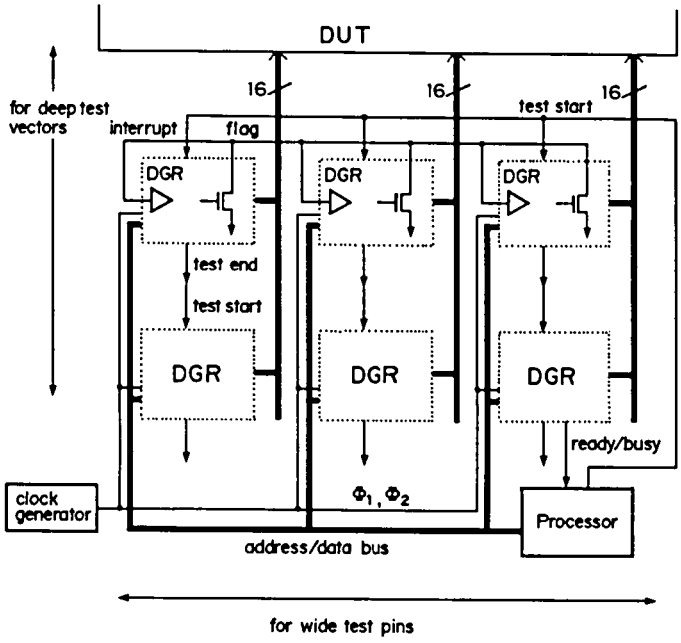


Fig. 5.1 A simple tester built using DGR chips.

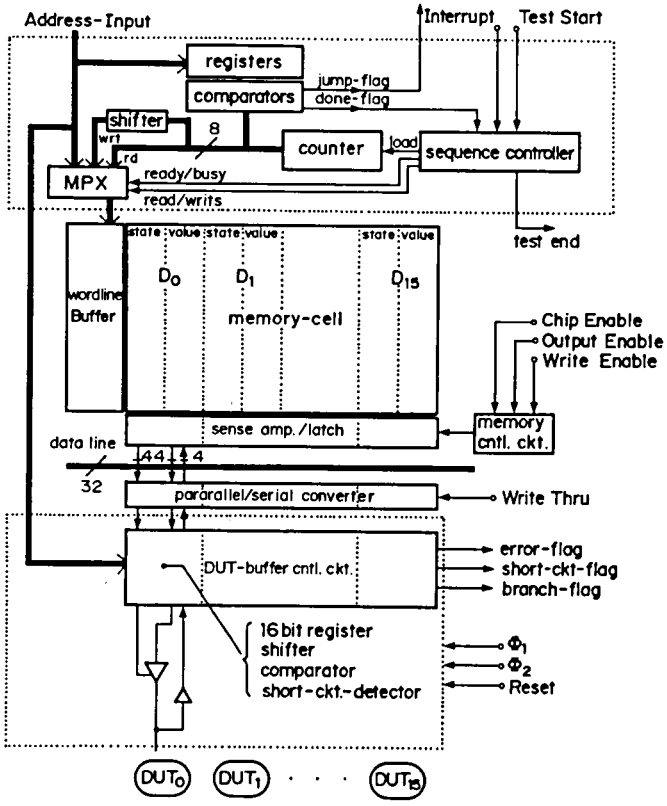


Fig. 5.2 A block diagram of the DGR.

configure the pin. The features of the pin drive are shown in Table 5.2. Each pin can be programmed for "Return Zero" (RZ), or a "Non Return Zero" (NRZ) code. The output timing can be synchronized to either the external Φ_1, or Φ_2 clock and either clock can be used to sample the input data. The state of the pin before the test begins is also user programmable, as is the state of the pin after the test finishes. For the testing of high impedance (HiZ) or open-collector device pins, DUT pins can be weakly clamped high or low. These clamps are implemented using a small PMOS and NMOS transistor in the DUT output buffer. The gates of these transistors are controlled by the configuration register.

Each pin is short circuit protected. If a short circuit condition is detected, the pad is tristated, and the short-circuit flag is asserted. The detail of this circuitry will be described in Section 5.4. There are two flags associated with error reporting. The "Error Enable Flag" causes the pin to compare the data on the DUT pad with the data supplied from the vector memory. If the values are not the same, the "Error Detected" bit is set in the configuration register, and DGR asserts Error to inform the host processor that an error has occurred. The remaining configuration bits are used to implement a simple conditional branch. If the "Compare Enable" bit is set, then the DUT pad is compared against the value stored in the "Compare 0/1" bit. When all the DUT pins with the Compare Enable set match their expected values a match signal is sent to the sequencer.

5.2.3 Sequencer

The sequencer consists of a counter, comparator, and a set of special registers. The functions of the sixteen special registers are listed in Table 5.3. Seven of registers are read-only and are automatically reset at the beginning of a test cycle. These include four registers to store the address of the first four errors, a register to record where the first pin short-circuit occurred, and then a number of registers for monitoring the test. The remaining nine registers are used to control the test sequence. The test begins at the "Start Address", and ends when it reaches either of the two "Stop Addresses". The "Jump Address" and the "Jump Destination" allow the user to program an unconditional loop, and the "Branch Address" and the "Branch Destination" provide a conditional branch capability. The "Trigger Address" and "Trigger Mask" allow the user to set up an external pulse synchronized to the tester. The mask selects which bits of the address must match the Trigger Address for the Trigger Pin to be high. This feature used in conjunction with the unconditional loop makes oscilloscope testing very easy. As we will describe in the next section, four vectors are fetched from the memory on each access. This improves the performance of the DGR chip, but also constraints the Start, Stop, Jump, and Branch addresses to be a multiple of four.

There are two registers which allow the user to configure the DGR as is shown in Tab. 5.4. The user can disable the jumps and branches by setting the appropriate

Tab. 5.2 DUT control registers.
A-reg.:even B-reg.:odd.

Data Out	Sync. Φ_1/Φ_2	A-0
	NRZ/RZ	A-1
	Start 0/1	A-2
	Start recv./drv.	A-3
	Stop Hz	A-7
	Invert Output	B-2
Pin Clamp	Weak Clamp 1	B-3
	Weak Clamp 0	A-6
Data Out	Sense Φ_1/Φ_2	B-0
	Current Pin Value	B-1
Flag	Short CKT. Enable	A-4
	Error Flag Enable	B-4
	Compare Enable	B-7
	Compare 0/1	B-6
Status	Error Detected	B-5
	Short CKT Detected	A-5

Tab. 5.3 Address registers.

Test Seq. CNTL. (R/W)	Start Address	120
	Jump Destination	121
	Branch Destinaton	122
	Stop Address 1	123
	Stop Address 2	124
	Jump Address	125
	Branch Address	126
Ext. Trigger (R/W)	Trigger Address	127
	Trigger Enable	128
Monitor (R)	Final Addres	129
	Counter Monitor	12A
	SC Occurrence	12B
	1st Error Occurrence	12C
	2nd Error Occurrence	12D
	3rd Error Occurrence	12E
	4th Error Occurrence	12F

Tab. 5.4 Status registers.

Status Reg.1 (130)	0	Jump Out Enable	R/W
	1	Branch Out Enable	
	2	Done Out Enable	
	3	Test Mode	
	4	Jump Input Acceptable	
	5	Branch Input Acceptable	
	6	MWB Enable	
	7	Jump Input Active Low	
Status Reg. 2 (131)	0	Jump Occurred	R
	1	Branch Occurred	
	2	Pin Initialize	R/W
	3	Done Generated	R
	4	Jump Flag Received	
	5	Branch Flag Received	
	6	Short CKT. Detected	
	7	Error Detected	

bits, and can set the memory protect mode by setting memory writeback enable (MWB Enable) low. The Pin Initialize bit is used to set the transition between the end of one test and the beginning of the next test. Until this bit is set the DUT pads will be holding the value from the last test. Once the bit is set the DUT pads are controlled by the initial value bits of the configuration register. As the pin values can be read directly as a bit of registers, this feature makes it possible to test the DUT chip step by step (Single Step Mode).

The Test Mode bit was added to make testing of the DGR chip, itself, simpler. Usually, the DGR chip only compares the data on the pad with the expected data for pins acting as receivers because of the possible conflict when the output is driven off a different clock than the one used to sample the input. Setting the Test Mode bit enables this check for drivers as well as receivers. This simplifies speed self-testing the DGR, since the DGR will check its own output.

All the registers are set to zero, when the "Reset Pin" is asserted. The registers and memory can be read or written through a simple synchronous host interface when the DGR chip is not running a test. In this state, the test vector clocks, Φ_1, and Φ_2 have no effect and can be ignored. When a test is in progress the host can still asynchronously read any of the registers, but register writes and memory accesss are inhibited.

5.3 On-Chip Memory Design

The operation of the vector memory is made more difficult by the dual use of the value bits. When "Memory Write Back" is enabled the value bits need to be read from and written to the memory on each cycle. A straight-forward implementation would require the memory access time to be half the vector cycle time. To prevent this situation, the DGR reads four vectors from the memory on each access, and stores four vectors on each write. Since each vector consists of two bits per pin and there are sixteen pins, the memory must read or write 128 bits on each access. By using this wide path to memory, the access time for the memory becomes twice the vector cycle time.

Even with the memory running at half the vector frequency, we felt that the memory access time would still affect the overall speed of the chip. To reduce this delay, we used a high speed memory organization, and took advantage of the synchronous nature of the vector access. The basic memory organization is shown in Fig. 5.3. The memory uses a standard six-transistor memory cell, with a sense amplifier and write driver allocated to each pair of bit lines. The memory has two modes of operation. During testing the memory accesses are synchronous, while during loading the host accesses are asynchronous. Figure 5.4 shows the timing waveforms for the synchronous access mode, along with the values of the control lines for asynchronous operation. To improve the access time for the synchronous operation, Φ_1 of the first clock cycle is used to equalize

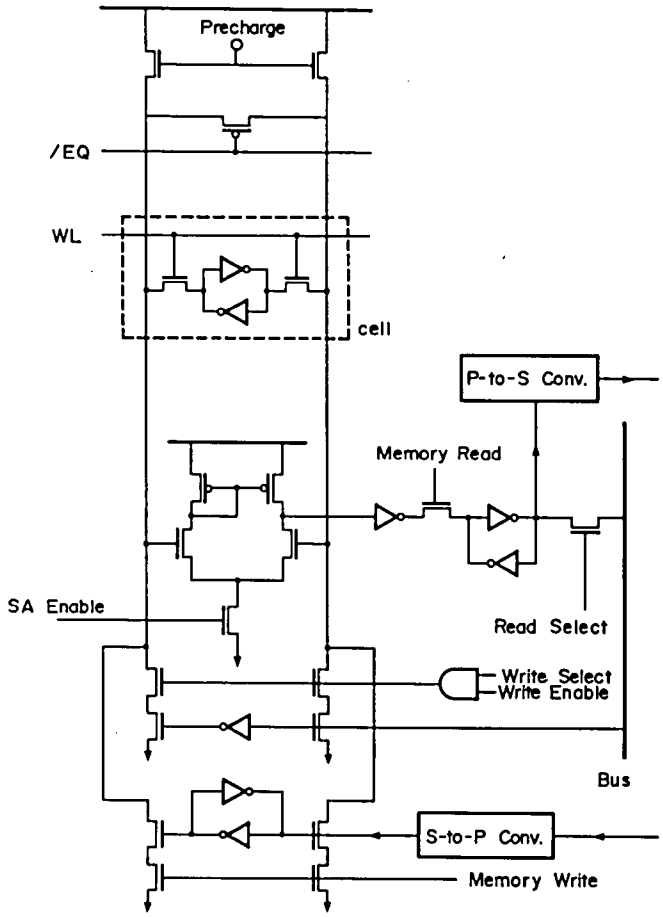


Fig. 5.3 Schematic of the memory showing peripheral circuits.

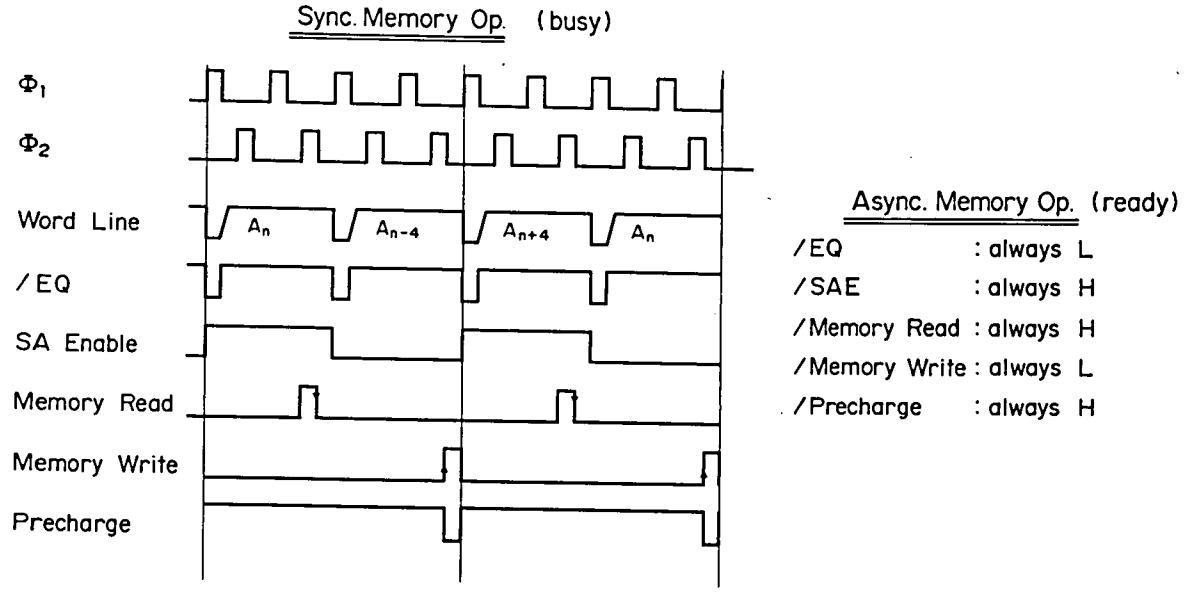


Fig. 5.4 Timing waveforms for the memory during synchronous operation.

the bit lines. During this period no word lines are asserted to prevent any write disturb problem. After the equalization the word line is driven into the array and the outputs of the sense amplifiers are latched on Φ_2 of the second cycle. To save power the sense amplifiers are only enabled during the read operations.

Figure 5.5 gives the timing waveforms during normal test operation. The basic memory operation takes four clock cycles, the first two clocks are used to read the memory, and the second two clocks are used to write the measured results back into the memory. To accomodate delays in DUT pin drive, the measured values are written into the memory 6 cycles after the expected value are read from the memory. The read address is presented to the memory early in the first clock cycle, and the output of the memory is latched at the end of the second clock cycle. The data is then serialized (Serial Out in Fig. 5.5) and sent to the DUT pin drive where it is first used on the fourth clock cycle. The measured values are collected during cycles four through seven, shifted into a serial to parallel converter ("Serial In" in Fig. 5.5) and then written into the RAM at the end of the eighth clock cycle. There is no problem with a write disturb since all reads equalize the bit lines before the word lines are asserted.

The comparison for branches and errors occurs four cycles after the address is first presented to the memroy. This skew causes branches to have a delay effect. If A_n is a conditional branch address, the DGR chip drives the Jump

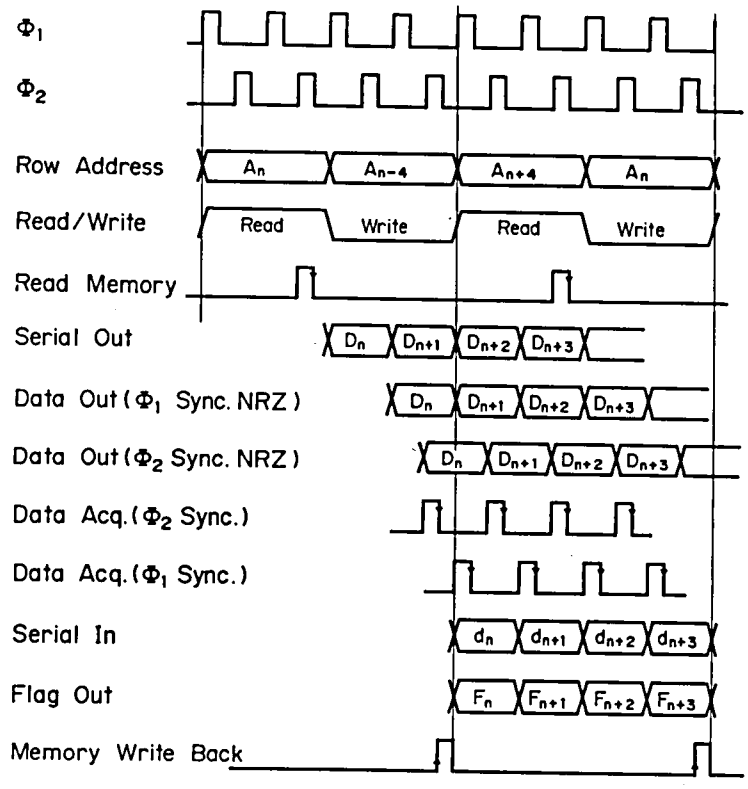


Fig. 5.5 Typical measurement sequence showing the parallel memory access and the parallel to serial conversion.

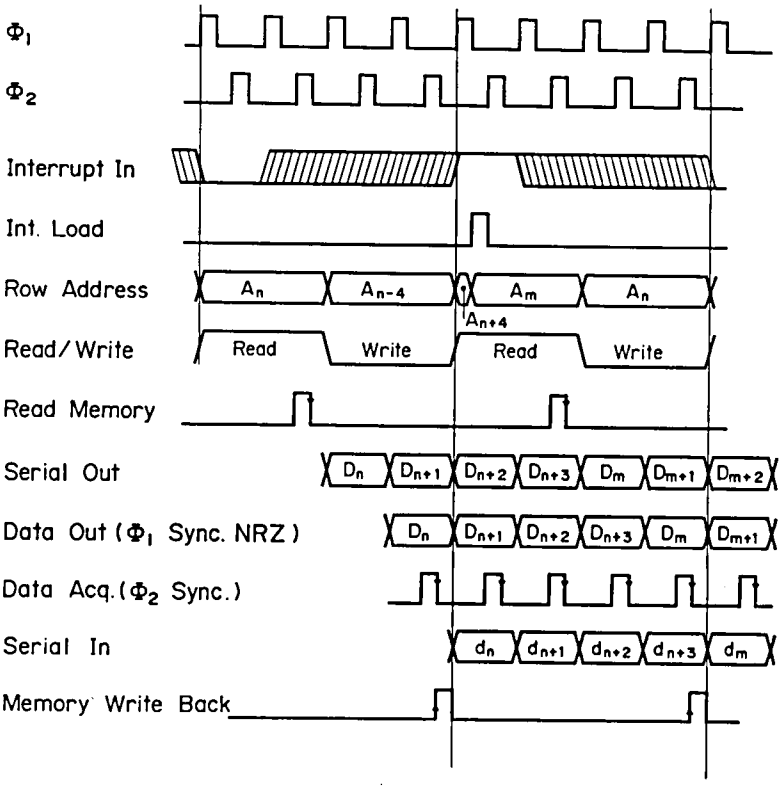


Fig. 5.6 Timing waveforms for Jump and Branch operation.

Pin when address A_{n+4} would be fetched. This means that vectors A_{n+1}, A_{n+2}, and A_{n+3} are driven out to the pads independent of the branch comparison. If data in all the DGR chips are correct then Jump will be high and this will cause the branch destination to be fetched instead of A_{n+4} (See Fig. 5.6). To provide some extra time for the external Jump Pin to settle, the loading of the address register ("Int. Load" in Fig. 5.6) is done in the gap between Φ_1 and Φ_2, rather than requiring the pin to settle early in Φ_1. This delay is acceptable since the memory word lines are not driven until Φ_2 of the first cycle.

The test begins when the "Begin Test" is asserted. The first clock cycle after "Begin Test" rises is used to set the counter and clear all the read-only address resisters. On the following clock cycle the start address is sent to the RAM to fetch the first set of four vectors. The first write into memory is not enabled since there is no valid data to store. The DGR continues to fetch values until the stop address is reached. At this point "Test End" pin is asserted, and it remains asserted until the "Begin Test" pin falls. The "Test End" pin can be used as the "Begin Test" signal to another DGR allowing two devices to be cascaded to increase the vector length. Besides this automatic stop by the DGR itself, the DGR vector output/input operation can halt by making "Begin Test" signal low by host. It enables emergency stop by the external controller.

To cover both synchronous and asynchronous operation,

the normally-on bit-line precharge scheme was used for a vector memory, which has already been addressed in Chapter 3. A smaller bit-line swing results in the higher speed access time (or, better write recovery), but consumes more current as shown in Fig. 5.7. By compromising the speed with the power, 1.5V was chosen as a bit-line swing. It is to be noted that the bit-line low level, 2V, is much higher than the level of flipping the cell data. It is quite important, if the wordline multi-selection occurs. Besides the current through the cell, there is another major DC current path, caused by the current mirror bit-line sense amplifiers. Different from the SRAM design in Chapter 3, 128 sense amplifiers are driven to be in active in the memory read phase of testing mode, simultaneously. So, the current dissipation per one sense amplifier was reduced and designed as 140uA.

Additional current consumes when one of bit-line pair is biased to zero in the memory write, It causes about 1.15mA current flow through the normally-on prechage transistor. As long as only 32 bit-line pairs are selected and usually the write enable input is short, it is not so significant in the asynchronous operation. However, in testing mode, especially in slower test, it might be a cause of heat-up, because 128 bit lines are set to zero. So, during the memory write of testing mode, the precharge load is switched to the higher impedance devices, and the power is saved.

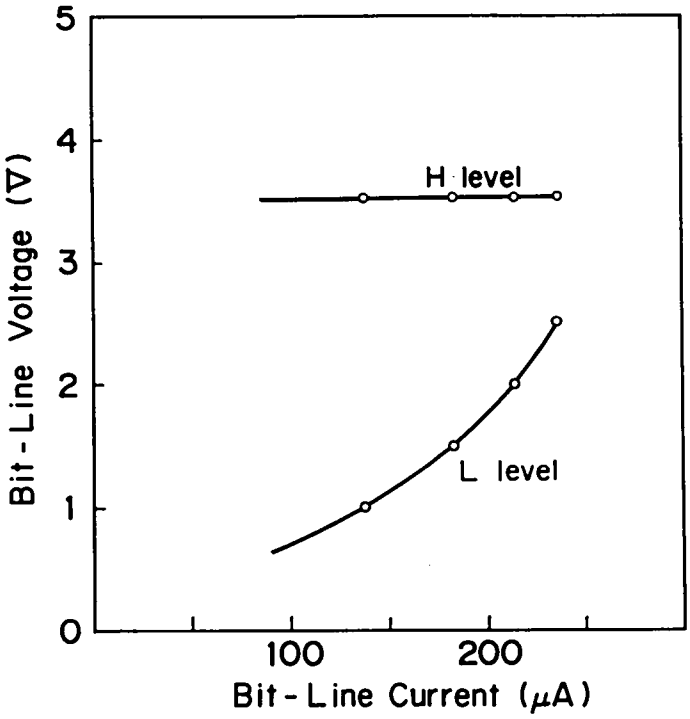
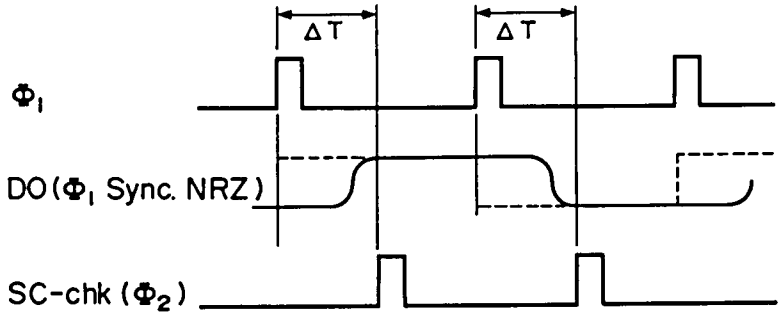


Fig. 5.7 Bit-line swing as a function of bit-line current.



ΔT : allowed settling time

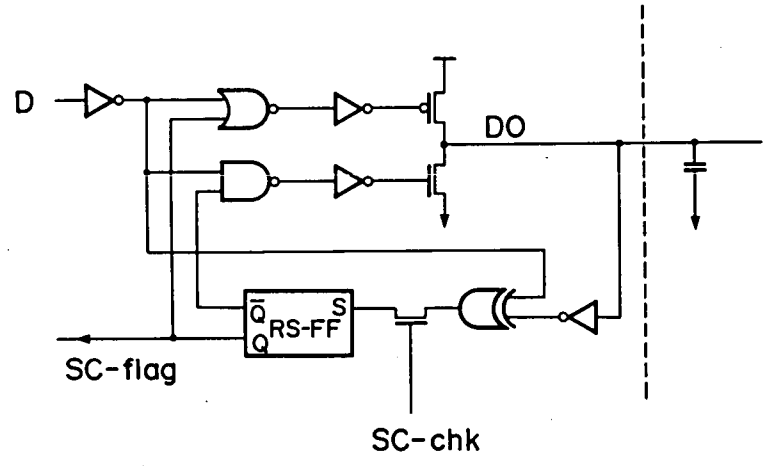


Fig. 5.8 Short circuit protection.

5.4 Circuit Descriptions

5.4.1 Short Circuit Protection

A Scheme of the short-circuit protection circuit, and a set of example waveforms are shown in Fig. 5.8. When the DUT acts as a driver, the data that drives the buffer and the data which actually comes out at the DUT pin (DO) are compared when "SC-chk" is high. The timing of "SC-chk" depends on the timing of the driver. If the pin is driven during Φ_1 then the check is done on Φ_2 to allow the pad enough time to settle to its new value. If the values do not agree then the short-circuit flag is asserted and that driver is disabled. This circuit will protect the outputs from melting from over-current. The short circuit flag can be read as part of the device configuration register and, if the short-circuit enable bit is set, will cause the address where the problem occurred to be reported.

5.4.2 Variable Data Acquisition

The skew between the clocks and the actual transitions of the DUT pads (which is defined Δt_0 in Fig. 5.9) is not significant as long as all the pins have roughly the same delay. This delay is not important since the DUT's clocks are also driven by the DGR chip (using a RZ format) and thus will be skewed along with every thing else. As long as all the pins are delayed the same amount the DUT can not

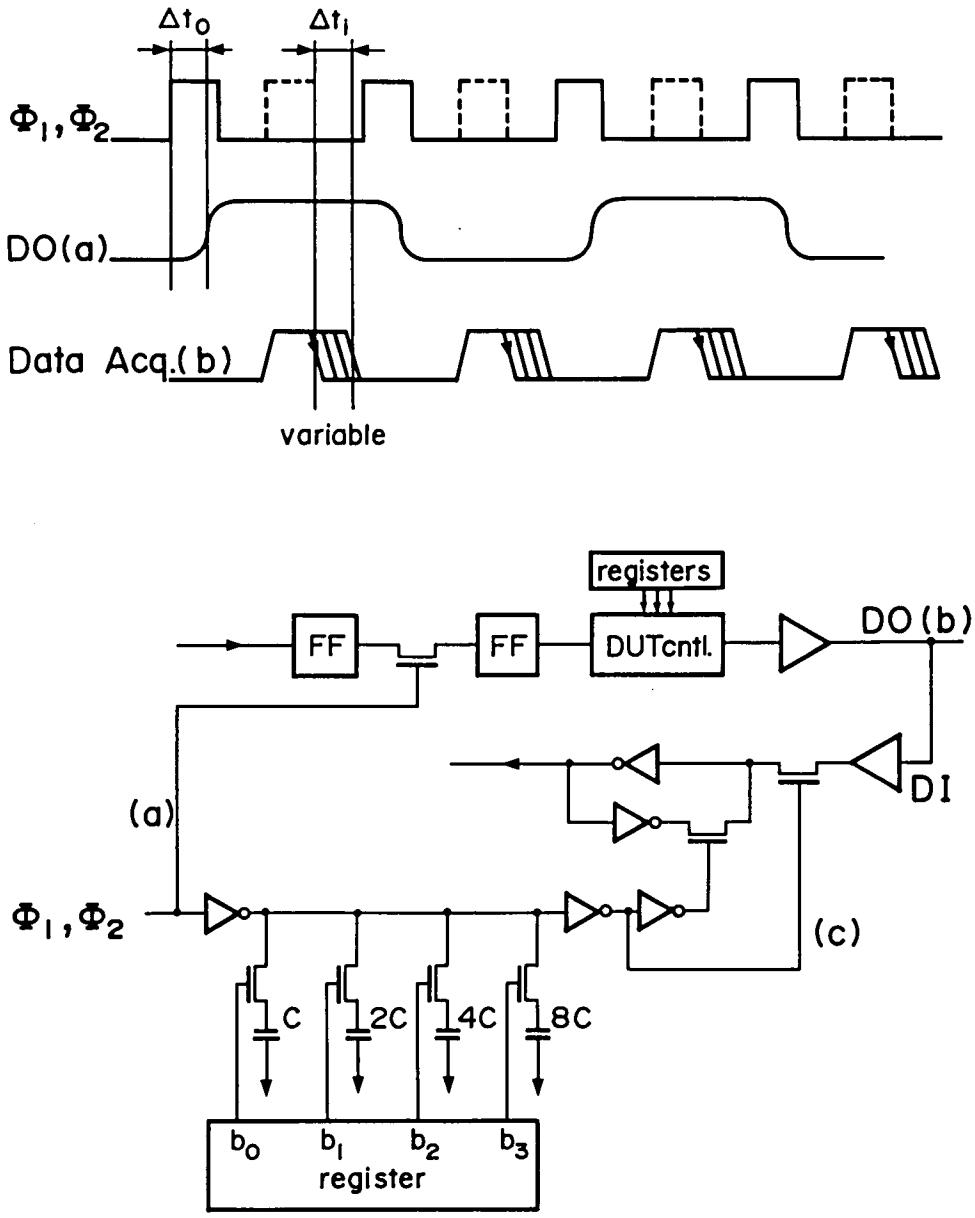


Fig. 5.9 Schemation of the adjustable delay circuit.

tell the skew is present. However, it would be more helpful to be able to adjust the skew between the clock and the input sample clock (Δt_i in Fig. 5.9) to insure that the DGR sampled the pins at the correct time. As shown in Fig. 5.9, the signal that drives the output buffer goes through many gates to provide an adjustable format, while the input path is quite simple. Using the switch level simulator RSIM [48], the output delay around 40ns and the input delay is 18ns. This delay is also dependent on the fabrication process which makes matching more difficult. To reduce this problem, we added a variable delay circuits to input clock. A set of binary capacitances and switches are used as a delay generator. A special 4-bit register controls the switches allowing the delay to be increased by increasing the load on the output.

5.4.3 Counter

As the DGR should operate from DC to high frequency vector rate, the address generator, or the counter was designed by static circuit with high-speed capability. A counter can be regarded as an adder. Now, we define that A and B are the adder inputs, C is the carry, and S is the sum outputs. Using the conception of the Manchester carry chain, the carry of the i-th stage, C_i, and the sum, S_i, may be expressed as;

$$C_i = G_i + P_i C_{i-1}, \qquad (5.1)$$

$$S_i = C_{i-1} \oplus P_i \qquad (5.2)$$

where

$$G_i = A_i B_i \quad \text{generate signal} \qquad (5.3)$$

$$P_i = A_i \oplus B_i \quad \text{propagate signal.} \qquad (5.4)$$

The notation $\oplus$ represents Exclusive OR. The adder elemental circuit is shown in Fig. 5.10(a). Especially, as a counter, $B_0 = 1$, and $B_i = 0$ $(i>0)$. So, Eqs (5.1) and (5.2) can be simplified as

$$G_0 = A_0, \text{ and } P_0 = A_0, \qquad (5.5)$$

$$G_i = 0, \text{ and } P_i = A_i \quad (i>0). \qquad (5.6)$$

By using the adder element in static, Exclusive-Or gates, made by 6 transistors shown in Fig. 5.10(b), 2-phase clocks, Φ_1 and Φ_2, and flip-flops, a static counter shown in Fig 5.11 was realized and applied to the address generator. It counts up the vector memory address one by one. When the jump flag is asserted, the content of the register, S^{i-1}, is replaced by the destination address, S^j, during both Φ_1 and Φ_2 low. In the following cycle, as soon as Φ_1 goes high, the address S^j comes out as the next address A_i.

Exclusive OR gates shown in Fig. 5.10(b) are also used as address comparators. This counter made by 3um CMOS technology, operates over 16MHz under 5V supply voltage, which is much faster than test vector generating rate, as explained in Section 5.6. So, it is concluded that the limiting factor of the maximum test vector rate is not the counter but the memory access time.

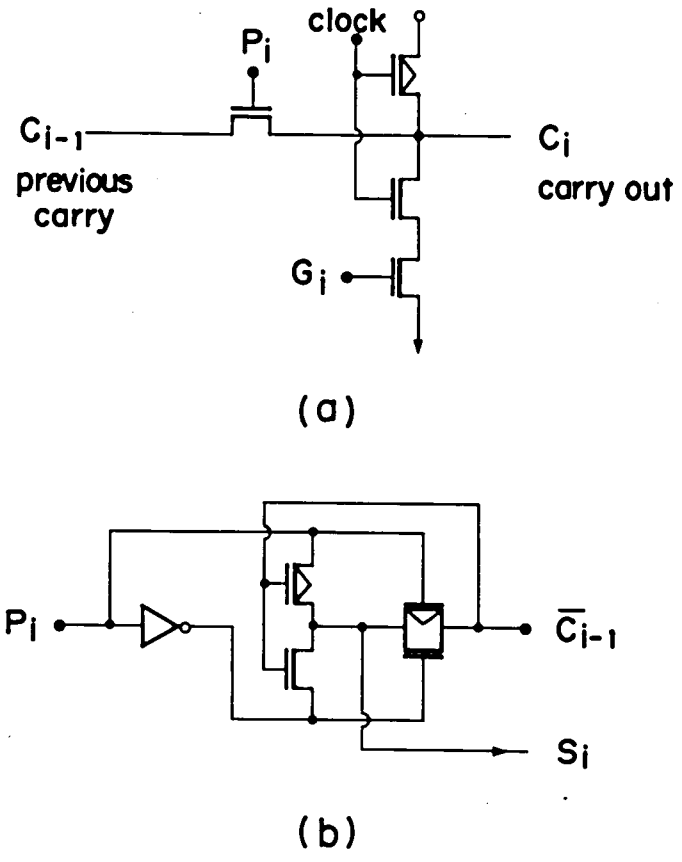


Fig. 5.10 Elemental adder circuits.

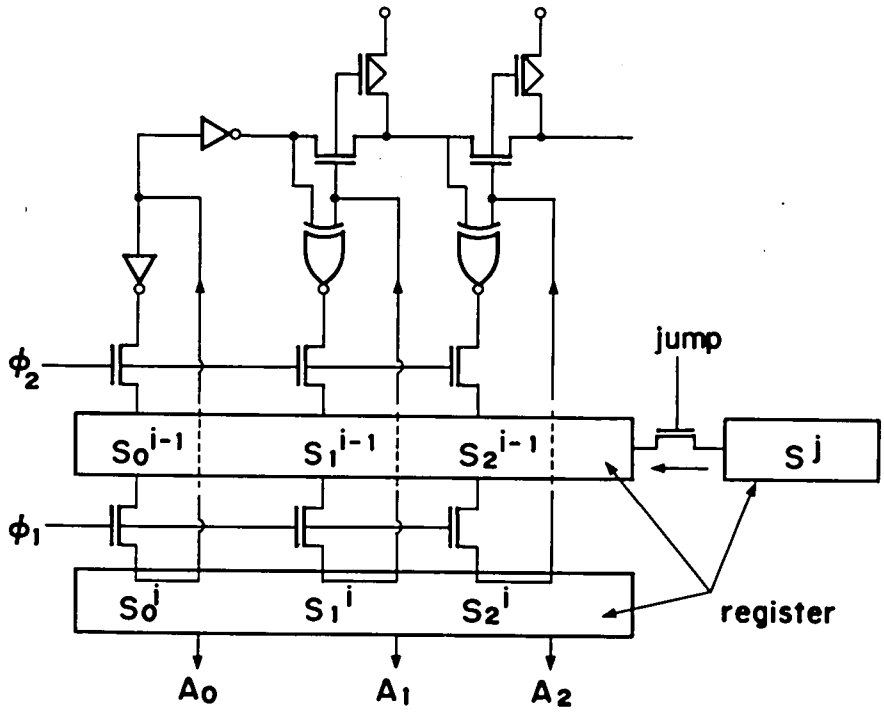


Fig. 5.11 Address generator.

5.5 Design Environment

Apart from the DGR itself, this section will describe the custom LSI design environment at Stanford University in 1985-86, where the DGR was designed.

The CMOS design rules applied to the DGR are shown in Table 5.5. Different from the standard memory design approach explained in Chapter 2, 3 and 4, all the design rules are described as unit λ. If 3um CMOS technology is taken, λ is 1.5um, while, with the same chip layout, λ becomes 1.0um, if 2.0um CMOS technology is applied. This approach loses chip-size optimization, but can keep the design property. This aspect seems to be more impotant than the individual design rule shrinkage from the designer point of view. And this is why the DGR was made by both 2um and 3um CMOS fabrication processes as presented in Section 5.6. The layout work of DGR was performed by "Magic", running on the micro VAX. "Magic" is an interactive pattern editting system for creating and modifying VLSI circuits [49]. It is based on the Mead-Conway style of design [50], and permits only Manhattan designs (those whose edges are vertical or horizontal). This means that Magic takes simplified design rules and circuit structure. The simplification makes it easier to design circuits and permits Magic to provide more powerful assistance.

In advance of this layout work, logic circuits was designed after the specifications were determined. In the stage, the Valid SCALD system, a schematics entry system was used. The graphic editor of the system has basic primitives like wires and circles, and also allows the user

Tab. 5.5 Design rules.

Width	
diff	2
poly	2
metal1	3
metal2	4
cut	2
via	3

Extension	
diff beyond trans	2
poly beyond trans	2
diff beyond cut	2
poly beyond cut	1
metal1 beyond cut	1
metal1 beyond via	1
metal2 beyond via	1

Separation	
same diff	3
ndiff-pdiff	10
poly-poly	2
polycontact-poly	3
diff-poly	1
metal1-metal1	3
metal2-metal2	4
cut-transistor	2
cut-via	3

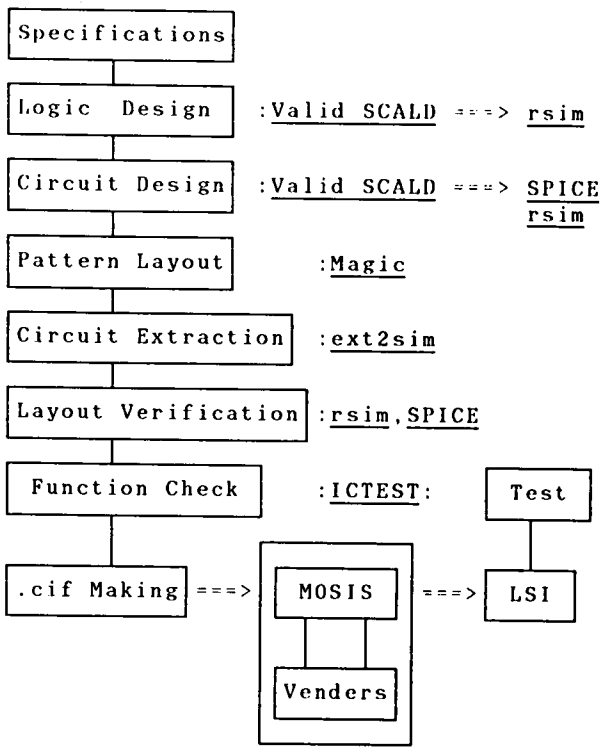


Fig. 5.12 CAD tools.

to call upon libraries which contain "bodies" that have been defined by the user. The bodies can have associated with the drawings, simulation and timing models. The SCALD system permits the compilation of a drawing or set of drawings so that the circuit can be logically simulated, checked for its timing or have a wire list generated.

Once the circuit schematic or the circuit layout is made, the net list is extracted and the form is converted to the "rsim" [48] input form. "rsim" is the event-driven, logic-level timing simulator. It adopts a linear model of the transistor in terms of an effective resistance of the transistor. The outputs are provided by the logic state, "0", "1", and "X", and the gate delay is calculated by RC network. The "rsim" solves the whole DGR circuits' state (77K transistor) in a reasonable time (1 step/sec). The simulator, "SPICE" was only used for the analog circuits such as sense amplifiers.

The whole circuits operation was verified by "ICTEST". "ICTEST" is a superset of the C programming language that specializes in test programs for integrated circuits. It is used by the designer to specify the stimulus to a circuit and the expected response from the circuit. In addition, "ICTEST" provides a common interface to a medium tester and simulation environment, for example "rsim". It means that by "ICTEST", one can make both the simulator input form for checking the function and the test vector for the real chip from the same source program.

Finally, the layout file is sent to MOSIS in ".cif" format, and the chip is fabricated by various vendors.

The design environment is summarized in Fig. 5.12.

5.6 Results

A prototype DGR was designed in a 3um, double Al CMOS technology, and contains 64.5K transistors in a die size of 9.2mm by 7.9mm. A die photo of device is shown in Fig. 5.13. One DGR supports 192 test vectors for 16 DUT pins, and fits into an 84 pin PGA package. The address access time using the asynchronous host interface as a function of the supply voltage is shown in Fig. 5.14. The difference between the outputs D0 and D8 is caused by the difference of the parasitic capacitance connected to the internal bus lines. Bus lines D0 through D7 connect together the outputs of all 51 registers in addition to the memory, while the internal buses from D8 to D31 only go to the memory. The diffusion capacitance from the additional bus drivers greatly increases the capacitance making the lower order lines much slower than the rest. Since the memory is read 128 bits at a time and there is a sense amplifier per bit, A_0 and A_1 are used only to select the correct set of latches to drive the internal data lines. Thus the difference in delay between the A_0 access time and the A_3 access time represents the time needed for word line selection and bit line sensing of the memory. The rest of the time is needed just to get the data off the chip.

In synchronous or vector mode, the minimum operating cycle time as a function of the power supply voltage is shown in Fig. 5.15. Figure 5.16(a) shows an oscilloscope photo of two DUT outputs, both running NRZ codes. One output is synchronized to Φ_1 and the other synchronized to

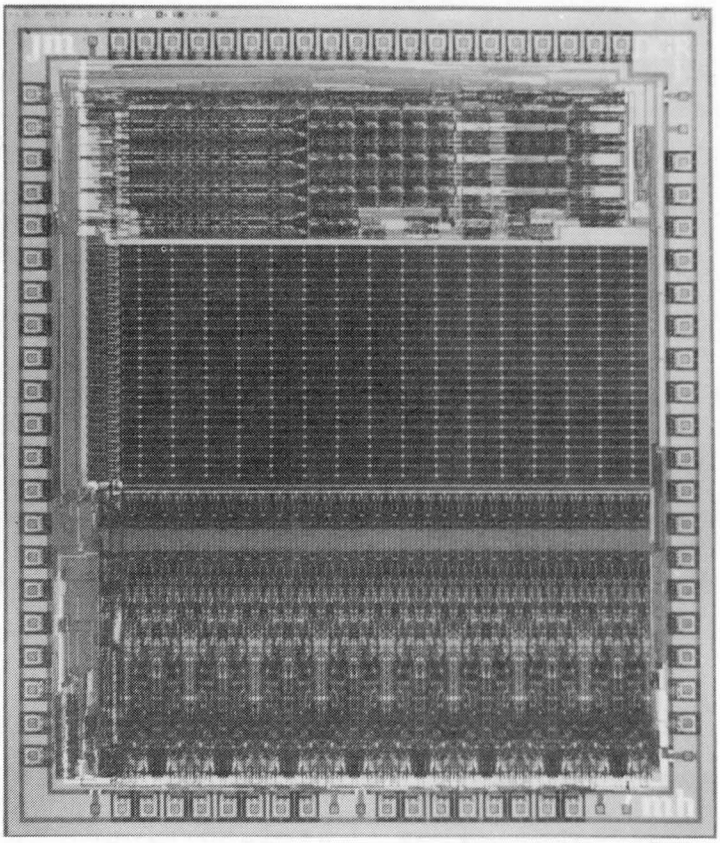


Fig. 5.13 Die photo of 3um DGR chip.

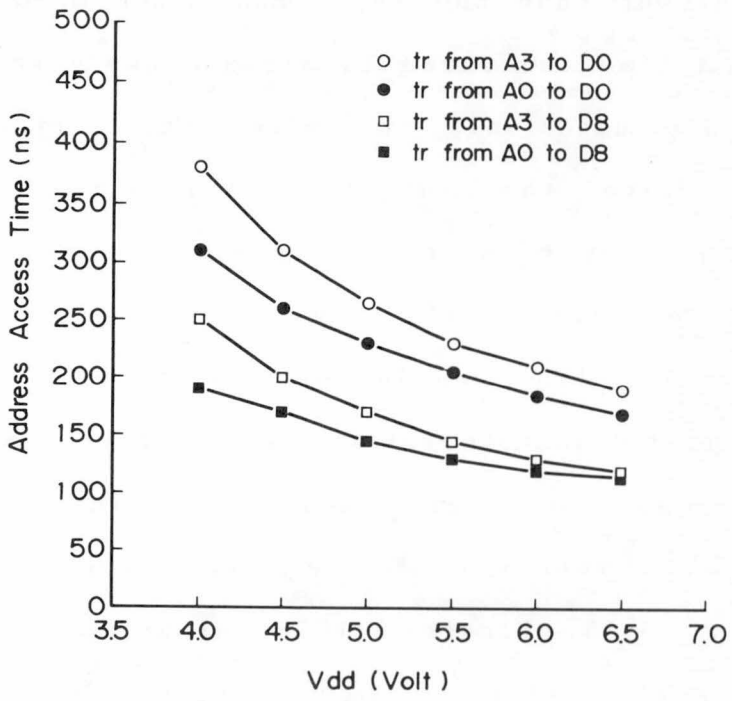


Fig. 5.14 Address access time as a function of supply voltage. The delay are given both for slow (D0) and fast (D8) data lines and full memory access (A3) and column select (A0).

Φ_2. This photo was obtained while the DGR was in a large test loop, counting from 0 to 192 and then jumping back to 0. Although the column access time in the asynchronous mode is over 150ns, a minimum operating cycle time of 90ns (11MHz) was obtained under the condition of Vdd = 5V. This result again shows most of the delay in asynchronous mode of operation is in driving the large internal data bus. The bus is isolated from the memory while the chip is in test mode. As shown in Fig. 5.16(a), there is a delay between the rising edges of the input clocks and the DUT output pulses. Although the skew between the external clocks and the pins is reasonably large, the skew between the pins is relatively small, less than 10ns on average. Figure 5.16(b) shows RZ and NRZ clocks both synchronized to Φ_2. Despite of the different generating path, the difference in the rising edges is kept within 10ns too. Figure 5.16(c) shows NRZ clocks synchronized to Φ_1 and a jump flag (JMP) detected at the output pin. It is spontaneously asserted as an interrupt for DGR and makes the test sequence change to the jump destination address. Over 10MHz operation is observed in those photos.

Figure 5.17 shows the high and low level output voltage as a function of the supply and sink current. The short-circuit detectable region is revealed in the same Figure. Once the short-circuit is detected, the DUT buffer goes to Hiz (High Impedance). So, the maximum current through the buffers is limited to 40mA as a supply current, and 45mA as a sink current, as long as the short-circuit

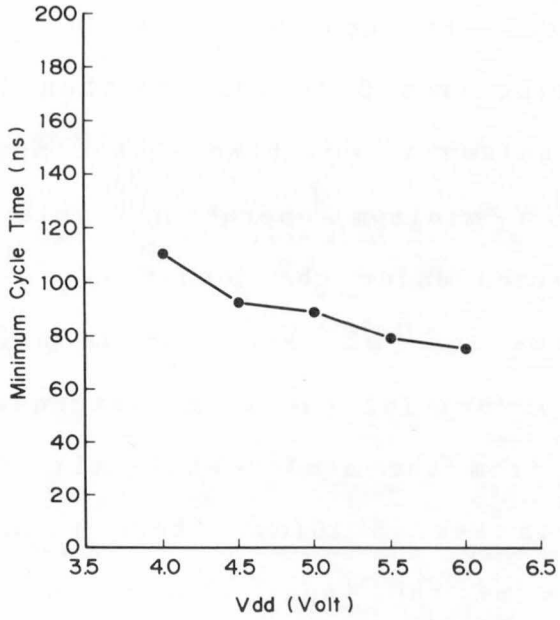


Fig. 5.15 Maximum vector rate as a function of supply voltage.

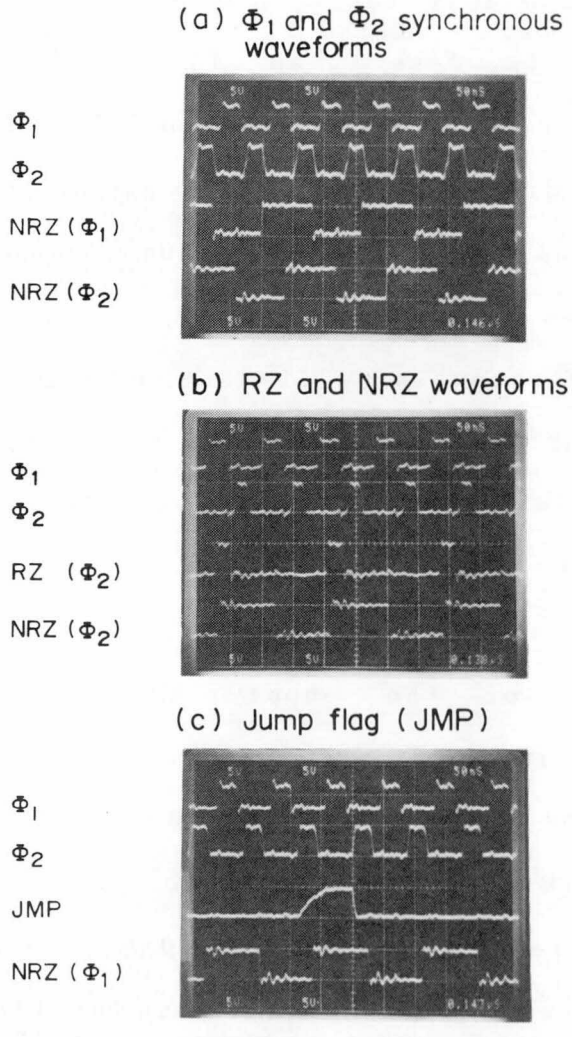


Fig. 5.16 DGR output waveforms of several working modes. Φ_1 and Φ_2 are the inputs. It is shown that DGR is working over 10MHz vector rate in every mode.

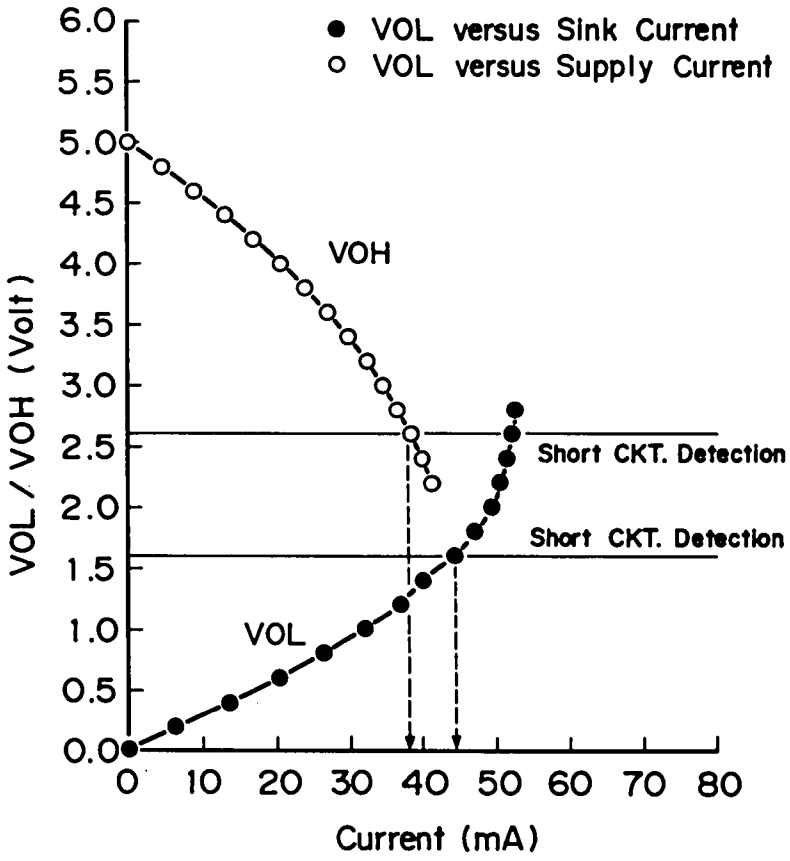


Fig. 5.17 VOH/VOL versus supply/sink current of DUT pin.

Tab. 5.6 Performance of 2um and 3um DGR.

item	2um DGR	3um DGR
Die Size	9.2*7.9mm	6.8*5.5mm
Transistors	65K	80K
DUT pin	16	16
Vectors	192	256
Vector rate	>10MHz	>16MHz
Idd	<65mA	<110mA

flag is enabled.

We have recently received a version of this DGR chip that was fabricated in a 2um technology. The chip is essentially a shrink of the original DGR chip with the vector memory extended to the full 256 vectors. Preliminary results indicated that this version of the chip runs over 16MHz, with the skews between the clocks and pins of roughly half the 3um version. Table 5.6 summarizes the results for the two chips.

5.7 Summary

We proposed an architecture for a single chip functional tester, the DGR, and have built a prototype of it. The chip contains a small vector memory, sequencer, and pin drive so it can drive and check DUT pins without needing to transfer data to the host at the vector rate. This allows one to build simple low cost testers that have a capability of short high-speed tests. The chips also provide a write through mode which makes it possible to exercise DUTs with a large number of test vectors. The prototype chips drive 16DUT pins and store 256 test vectors on a 6.8mm by 5.5mm die using a 2um CMOS technology. These chips can be connected in parallel to create more DUT pins, and cascaded to increase the vector length. The 2um versions of the chip operate over 16MVector/sec, and could easily be extended to a higher frequency and deeper vector

length by using a more advanced CMOS technology.

From an ASMIC point of view, on-chip memory is successfully embedded into the system. The sequencer points the memory address in advance, and each pin is able to use the memory data in an internal access time, in a word, a delay from word-line buffers to sense amplifiers. Different from the standard memory design, the following techniques were applied to meet the specifications.

4-word simultaneous read and write,
combining the pipeline scheme.

Co-existence of synchronous and asynchronous
memory access.

Wide 128bit parallel organization.

Unique power saving bit-line load.

All are necessary for the architecture to achieve the flexible and high-speed LSI testing.

144 項欠

Chapter 6

Conclusions

The device and circuit designs of an ECL-RAM, a Bi-CMOS RAM, an EEPROM, and an ASMIC for LSI function testing, containing CMOS vector memory in it, have been described. They have different bit density, and are classified as different categories. Among these memories, the ECL RAM is the fastest, but the memory cell consumes holding current to keep the data. Additionally, it occupies the largest area due to the bipolar transistor isolation. The CMOS six-transistor memory cell used for the Bi-CMOS RAM and the ASMIC holds the data without DC current, matching to the large scale integration. The cell holding data statically enables to achieve fast access time, when a bipolar circuit technology is applied. The stable cell can be accompanied by more flexible peripheral circuitry than the standard read/write scheme. In other words, the RAM cell combined with suitable read/write scheme for the application, enables a high-speed LSI functional tester, DGR (Data Generator and Receiver).

The RAM data is volatile, and occupies still larger cell area than non-volatile memory cell. An EEPROM cell is composed of three electric components, a floating-gate transistor, a select-gate transistor and a tunneling region made by thin oxide. However, it suffers from a milli-second order programming time, and slow access time due to the

small conductance cell structure. Moreover, the process complexity prevented it from integrating large bits in a chip.

This chapter will first summarize the characterstics of individual devices, and demonstrate their situations in memories. Next, the common probelms through the memory design, such as high-speed capability, high-functionality, and design methodology will be discussed. Finally, some thoughts will be addressed to conclude the thesis.

6.1 Individual Device Results

For the design of fast ECL RAM, high performance electric components were assembled, these are, oxide isolated NPN transistors, Schottky diode, and double level aluminum. Polysilicon resistor was newly applied to bipolar RAM. Utilizing minimum 3um design rule, the cell and chip size is 34*55um and 2.0*2.3mm, respectively. The 256bit RAM presents 4.6ns access time with 340mW power dissipation.

Collector isolated NPN transistors are newly adopted to CMOS circuits in the design of 64k high-speed RAM. The RAM uses 6-transistor memory cell, MoSi gate and bipolar bit-line sense amplifiers. The bipolar transistor not only decreases bit-line sensing delay, but also decreases the chip-to-chip variation, namely, the worst-case access time. The RAM offers 28ns address access time with 225mW power dissipation at 5V power supply. The design rules are basically 1.8um besides the NMOS gate length of 1.5um. NPN transistors were fabricated on a same chip without any additional process step other than CMOS. The cell size and chip size are 18*20um, and 5.95 * 6.84mm, respectively. The CMOS RAM access time has been greatly improved from the range of 60-80ns to around 30ns.

From the investigation shown in Chapter 3, the single-polysilicon cell is proved to be suitable for high-density EEPROM, if the layout pattern and device parameters are well optimized. The key issue for improving the performance is making the capacitance between the floating gate and control gate large by using thin oxide as

a di-electric. From the analysis, the tunneling effect is negligible, if the threshold shift is within 5V. However, within the recent available technology, the thin oxide area on a drain should be separated from the floating gate transistor because of the substrate current generation, which, however, is accompanied by a cell size increase. The newly developed 256Kbit EEPROM enables 150ns address access time with 80mW power dissipation. The programming time is improved to 225us/byte virtually. By using 1.4um design rule, the cell size is reduced to 6.5*11.5um, resulting in 7.33 * 6.23 mm chip size. It is contained into 28-pin DIP, which is pin-to-pin compatible to 256Kbit SRAM.

One chip tester chip, DGR, has a novel architecture, giving a solution to the high-speed testing for recent diversities of ASIC (Application Specific IC). In spite of the quite conservative 3um CMOS technology, the chip enables over 10MHz test vector generation, DUT data acquisition, and comparison with the expected data. The on-chip memory gains the operating band-width by 4 word parallel read/write scheme, and pipelining. The RAM is organized to fit 16 DUT pins, with the depth of 192 test vecters. The power dissipation is 325mW. Using 37*28um RAM cell, the chip size is 9.2 * 7.9 mm, 84-pin PGA packaging. The advanced DGR applied by 2um CMOS technology, having 8Kbit vector memory, offers over 16MHz vector rate with 550mW.

6.2 High-Speed Capability

High-speed memory access is required more than ever, as the development of high-speed processors reduces the bottle-neck of the system performance to the memory band-width. The standard memory access time is roughly divided into the following 5 components, input-buffer delay, decodeing, word-line driving, bit-line sensing, output-buffer delay. According to the MOSFETs' scaling law, device performance has been improved by the device miniaturiza-tion, which has made the memory bit capacity larger than the previous one. As long as the external circumstances, or the I/O specifications are not changed, the input and output buffer delay would be improved by device perform-ance. The decoding delay at least keeps constant, in spite of the increase in line numbers and total line length. However, nowadays, the assumption may not be true, because specifications which are not necessary at all for the previous generation, such as the noise immunity, electro static persistence, and supply current saving, become serious. The clearance sometimes sacrifices access time. So, the delay reduction in the cell peripheral has become more important than ever,where device miniaturization does not always lead to the delay reduction. The parasitic re-sistance and capacitance can not be scaled down as devices.

The word-line delay reduction is enabled by using low resistivity material. Silicide, such as Mosi, having 5Ω/□ resistivity as is applied to Bi-CMOS RAM, is one solution. However, it is only one sixth RC-delay reduction of poly-

silicon. Double level Aluminum, 30mΩ/□ resistivity, reduces RC delay more, as is used for DGR. However, it needs additional contact area, connecting the gate materal with metal, rusulting in core area increase, as shown in DGR and ECL design. After the RC delay reduction, a word-line buffer with large drivability should be necessary, for example, Darlington Configuration, and Bi-CMOS driver. Different from bipolar memory, word-line amplitude of CMOS memory should be large enough to obtain much cell current, I_{CELL}. It is common feature to all the voltage driven type devices as MOSFET, but is a disadvandage for achieving high-speed word-line switching. Because, word-line buffers should provide much current for charging the product of capacitance and amplitude. There is an optimization point for the power dissipation, the layout pattern, and the speed, as to a word-line buffer.

Bit-lines are always made by Aluminum, so no RC delay problem occurs. In case of SRAM, a pair of bit-lines vary differentially. Using the precharge currnt I_{PRCG}, and I_{CELL}, the bit-line delay, t_{BL}, is approximately represented by:

$$t_{BL} = C_{BL} * \Delta V_{BL} / (I_{CELL} + I_{PRCG}) \qquad (6.1)$$

where, C_{BL}, and ΔV_{BL} are bit-line capacitance and minimum detectable bit-line amplitude, respectively. Assuming the normal design as $I_{CELL} = I_{PRCG}$, t_{BL} is then,

$$t_{BL} = 0.5 * C_{BL} * \Delta V_{BL} / I_{CELL}. \qquad (6.2)$$

C_{BL} increases simply proportional to the square root of bit capacity, as it is composed of the cell diffusion region and metal lines. I_{CELL} is determined by the cell structure.

In case of ECL RAM, the limiting factor of I_{CELL}, or read current I_R was the parasitic resistance of Schottky diode. That of SRAM was the equivalent series resistance of the bit-line pass and cell driver NMOS transistors. As the gate length is determined by the minimum design rule, the I_{CELL} is inceased by the gate width, which, however, leads to the cell size increase. Consequently, reducing the bit-line amplitude, ΔV_{BL}, is effective to improve the delay without area penalty. Adopting a small signal detectable sense-amplifier, as bipolar differential amplifier which was used in Bi-CMOS SRAM design, is necessary. Bipolar transistors naturally do not need large amplitude input to get a drivability. It is proved to be effective.

Single ended bit-line as EEPROM or DRAM has a handicap in the bit-line delay in comparison with SRAM, because the reference voltage does not vary differentially, or $I_{REG}=0$. In the same ΔV_{BL} as SRAM, t_{BL} is two times larger at least. Precharging and equalizing bit-line pairis by internally generated clocks aid bit-lines to recover fast from the previous level, as far as the clock width and its drivability are optimized. It is the other approach to decrease t_{BL}. It is proved to be effective by the EEPROM design.

Other than the individual technology to reduce the bit- and word-line delay, dividing memory cell array into several blocks is useful, although it is accompnied by a little area increase. It is adopted to SRAM and EEPROM design. The architecture helps to save the active current consumption, as well.

6.3 High Functionality

Having high functionality on a chip has been attractive from the system designer's point of view, and this is comming true for LSI designers. Through the devices in the thesis, the functionality is adopted for improving the device performance. Different from ECL and SRAM design, there is a room for taking it in EEPROM and Single chip tester design.

The page-mode programming in EEPROM enables to shorten the virtual programming time. After the successive data loading to latches in nano-seconds, the data are programmed in parallel, taking advantage of the low current dissipation in programming. The same technique will be utilized for more than 1Mbit UV-EPROM. The page-mode programming can be performed by adding several glue logics externally. For users, the page-mode is the similar function as "gang", which programs several devices in parallel, implemented into some of PROM writers. However, for venders, they are different. The programming time reduction per device, even if it does virtually, cuts off test cost. So, it becomes necessary for the high-density non-volatile memories.

Dealing with the data in parallel is quite effective to gain the performance, as was shown in the tester chip design. Pipelining in 3um DGR enables below 90ns test vector generation, although the memory access time in worst case was 280ns at 5V. The technique is an extended idea from the RISC (Reduced Instruction Set Computer) type micro processor as MIPS-X [51]. The one of the greatest problems

of pipelining is a jumping, or changing the process sequence. DGR only allows an interupt input in every four clock cycles, to avoid NOP (no operation) cycle. MIPS-X avoids the same problem by the compilation from the source to assembler. Different from MIPS-X, changing the process execution or the NOP insertion are not allowed for DGR, because it needs a real time response.

As the bit capacity increases, the functionality will be required not only to eliminate some peripheral glue logics around the LSI but also to gain the performance of LSI itself, or to reduce the LSI cost.

6.4 Design Methodology

The design methodology difference between the four memories raises interesting discussions. The former three devices, ECL, SRAM and EEPROM, took the methods of standard memory design. As the layout was hand-crafted, it was well optimized. The design and fabrication were closely related and interacted each other. As a result, the detailed design rules were optimized to meet requirements of individual devices, even though the same "minimum 2um" design rule is applied. Special device structures, such as bipolar NPN transistor in CMOS, or the single-polysilicon structured cell are acceptable, if it is effective for the device performance. On the contrary, the turn-around-time is considerably longer, which means that it takes a long time to debug the circuit and system design. Moreover, it is too

difficult to shrink the chip at once. The approach is suitable for the mass-produced devices.

On the other hand, the design of the tester chip is quite different. No other process step than that for CMOS logic gates is permitted. As a result, the cell size used for the tester chip was about five times larger than that of Bi-CMOS RAM, although both took the same six transistor memory cell structure, and the approximately same 2um design rules were applied. Instead, because of its simplified design rules, the design was fully supported by CAD, and the design period per one transistor was reduced to one tenth of the standard memories. In addition, as the design property was kept independent of the fabrication process generation, the 2um device was immediately realized as an extention of 3um version. From the designer point of view, it would be more preferable to have a storter turn-around time, as was insisted in the panel discussion of 1985 IEDM among the designers. This is an important approach for the custom IC design.

6.5 Final Thoughts

Bi-CMOS is gaining a popularity, because of its high sensitivity and high drivability. Since then, a lot of Bi-CMOS circuits have been proposed not only for the high-speed memories but also for some logic LSIs. The bipolar amplifier application in CMOS circuits is expanded to a high-speed carry transfer circuit of the adder. The combi-

nation of oxide isolated shallow junction NPN transistors with buried layer and miniaturized CMOS transistors will make it possible to realize high performance LSI.

A single polysilicon cell is still attractive for the high density EEPROM in the logic combined systems, because of the fabrication process compatibility. From the bit density and the cell current point of view, a cell built with a tunneling area coinciding with the channel of the thin oxide transistor is attractive. The drain profile, oxide thickness, and charge pump circuits should be optimized for high performance EEPROM.

The concept of silicon tester is quite new. There will be several approaches to reduce the evaluation cost of ASIC. For the next generation one- chip tester, implementing internal pattern generator, dealing with three values ("0","1" and "X") assigned to one DUT, increasing the test vector depth rather than the width, and obtaining more timing-accuracy should be taken into consideration.

The bit density of every standard memories is getting large, almost 4 times every 3 years. The tendency has not been changed, although the process has become more sophisticated. On the other hand, we have seen an explosion in the development of function integrated chip as ASMIC (Application Specific Memory IC). The high-speed and high-functional CMOS memory has been demanded continuously. The innovative design approach will be required to find the break-through point toward the next generation, cooperating with the fabrication technology.

156 項欠

References

(1)H.Ikoma, K.Ichinose, K.Kanzaki, S.Shinozaki, K.Fujita, J.Miyamoto, Y.Miyasako, "Introduction of VLSI (4), Bipolar Integrated Circuits", 1984, Kindai-Kagaku-sha. (in Japanese)

(2)S.Shinozaki, T.Iizuka, F.Masuoka, K.Shinada, and J.Miyamoto,"Role of the External n-p-n Base Region on the Switching Speed of Integrated Injection Logic(IIL)", IEEE J. Solid-State Circuits, vol.SC-12, no.2,p.185,April 1977.

(3)S.Shinozaki, K.Shinada, and J.Miyamoto, "Effects of Gate Geometry on Propagation Delay of Integrated Injection Logic (IIL)", IEEE J. Solid-State Circuits, vol.SC-13, no.2, p225, April 1978.

(4)T.Sugano, T.Sakurai, H.Ohiwa, M.Sugawara, K.Natori, A.Kanuma, R.Dan, J.Miyamoto, S.Saitoh, "Introduction to MOS LSI Design", (translated in Japanese), 1983, Sangyou-Tosho.

(5)J.Miyamoto, "Recent Development and Application of 64K and 256K EEPROM", lecture text, Japan Technology Information Center, 1985. (in Japanese)

(6)J.Miyamoto, K.Shinada, S.Shinozaki, and N.Sekiguchi, "Application of Polycrystalline Silicon Load For High Performance Bipolar Memory", in IEDM (International Electron Devices Meeting) Tech., Dig., p50, Dec. 1980.

(7)J.Miyamoto, S.Saito, H.Momose, H.Shibata, K.Kanzaki, and S.Kohyama, "1.0 um N-Well CMOS/Bipolar Technology for VLSI Circuits", in IEDM Tech.,Dig., p.63, Dec. 1983.

(8)J.Miyamoto, S.Saito, H.Momose, H.Shibata, K.Kanzaki, and T.Iizuka, "A 28ns CMOS SRAM with Bipolar Sense Amplifiers", in International Solid-State Circuit Conference(ISSCC) Dig. Tech. Papers, p.224, Feb. 1984.

(9)J.Miyamoto, S.Saito, H.Momose, H.Shibata, K.Kanzaki, and T.Iizuka, "A High Speed 64K CMOS SRAM with Bipolar Sense Amplifiers" ,IEEE J. Solid-State Cicuits, vol.SC-19, no.5, p.557, Oct.,1984

(10)H.Momose, H.Shibata, S.Saito, J.Miyamoto, K.Kanzaki and S.Kohyama, "1.0 um N-well CMOS/Bipolar technology", IEEE, J. Solid-State Circuits, vol. SC-20, no.1, p.137, Feb. 1985.

(11)N.Matsukawa, S.Morita, K.Shinada, J.Miyamoto, J.Tsujimoto, T.Iizuka,and H.Nozawa, "A High Density Single Polysilicon Structure EEPROM with LB(Lower Barrier Hight) Oxide for VLSI's", in 5th,Simp. VLSI Technol. Dig., Tech. Papers, p100, May 1985.

(12)J.Tsujimoto, J.Miyamoto, N.Matsukawa, K.Shinada, S.Morita, H.Nozawa,and T.Iizuka, "A 5V-Only 256K CMOS EEPROM using Barrier Height Lowering Technique",in 11th European Solid-State Circuits Conf. Dig., Tech., p.241,Sept.1985.

(13)J.Miyamoto J.Tsujimoto, N.Matsukawa, K.Shinada,. S.Morita,H.Nozawa,and T.Iizuka, "An Experimental 5-V-Only 256-kbit CMOS EEPROM with a High-Performance Single-Polysilicon Cell",IEEE J. Solid-State Circuits, vol.SC-21, no.5, p.852, Oct.,1986.

(14)J.Miyamoto and M.A.Horowitz,"A Single-Chip Functional Tester", in ISSCC Dig. Tech. Papers, p.232, Feb., 1987.

(15)J.Miyamoto and M.A.Horowitz, "A Single-Chip LSI High-Speed Functional Tester", IEEE J. Solid-State Circuits, vol.SC-22, no.5, p.820, Oct., 1987.

(16)S.Konishi, J.Masuhara, T.Ohtani, M.Sekine, M.Isobe, T.Iizuka, Y.Uchida, and S.Kohyama,"A 64K CMOS RAM", in ISSCC Dig. Tech. Papers, p.258, Feb.1982.

(17)K.Ochii, K.Hashimoto, H.Yasuda, M.Masuda, H.Nozawa, and S.Kohyama,"A 15nW stand-by power 64K CMOS RAM", in ISSCC Dig. of Tech. Papers, p.260, Feb. 1982.

(18)O.Minato, T.Masuhara, T.Sakai, Y.Sakai, and T.Hayashida, "A Hi-CMOSII 8Kx8 bit static RAM",in ISSCC Dig. Tech. Papers, p.256, Feb.1982.

(19)K.H.Hardee and R.Sud, "A fault tolerant 30ns/375mW 16Kx1 NMOS Static RAM", IEEE J. Solid-State Circuits, vol.SC-16, p.435, Oct.1981.

(20)A.V.Evel, G.E.Atwood, E.Y.So, S.S.Liu, V.N.Kynett, R.M.Jecmen, J.Mingo, and H.Dun, "An NMOS 64K static RAM", in ISSCC Dig. Tech. Papers, Feb. 1982.

(21)K.Toyoda, M.Tanaka, H.Isogai, C.Ono, Y.Kawabe, and H.Goto, "A 15ns 16K ECL RAM with a p-n-p load cell", in ISSCC Dig. Tech. Papers, Feb.1983.

(22)Y.Kato, M.Odaka, K.Ogiue, H.Miwa, and K.Matsumura, "A 16ns 16K bipolar RAM" in ISSCC Dig. Tech. Papers, Feb.1983.

(23)T.Masuhara, S.Minato, Y.Sakai,and M.Kubo, "A high-speed, low power HiCMOS 4K static RAM", in ISSCC Dig. Tech. Papers, p.108 1978.

(24)L.H.Edwin and L.S.Stephen,"An ECL compatible 4K CMOS RAM", in ISSCC Dig. Tech. Papers, p.248, Feb.1982.

(25)K.Ogiue, M.Odaka, S.Miyaoka, I.Masuda,T.Ikeda, K.Tonomura, and T.Ohba, "A 13ns/500mW 64Kb ECL RAM", in ISSCC Dig. Tech. Papers, p.212, Feb.1986.

(26)P.Hickman, F.Ormerod, and D.Schucker, "A High Performance 6000 Gate BIMOS Logic Array", in Proceedings of the Custom Integrated Circuits Conference (CICC), p562, May, 1986.

(27)P.i.Suciu, M.Briner, C.S.Bill, and D.Rinerson, "A 64KEEPROM with extended temperature and page mode operation", in ISSCC Dig. Tech. Papers, p.170, Feb.1985.

(28)R.Jolly, R.Tesch, K.Campbell, D.Tennant, J.Olund, B.Cremen, R.Lefferts, and P.Andrews, "Two 35ns 64K CMOS EEPROMs", in ISSCC Dig. Tech. Papers p.172, Feb. 1985.

(29)S.Mihrotra, T.C.Wu, T.L.Chiu, and G.Perlegos, "A 64Kb CMOS EEPROM with on-chip ECC", in ISSCC Dig. Tech. Papers p.149 Feb. 1984.

(30)F.Jones, and A.Lancaster, "EEPROM adapts easily to in-system changes", Electronic Design., Aug. 1983, p189.

(31)C.Kuo, J.Yeargain, W.downey, K.Ilgensten, J.Jorvig, S.Smith, and A.Bormann, "An 80ns 32K EEPROM using the FETMOS cell", IEEE J. Solid-State Circuits, vol.SC-17, no.5, p.821, Oct.1982.

(32)G.Landers, "5-volt-only EEPROM mimics static RAM timing.", Electronics, vol.55, p.127, 1982.

(33)S.Logie, E.Harari, S.Li, W.Liu, and R.Das, "A new floating gate cell and technology for a 5V only CMOS 16K EEPROM", in IEDM Dig. Tech. Papers ,Dec. 1983.

(34)M.Wada, M.Asano, H.Iwahashi, S.Inoue, R.Kirisawa, K.Hieda, and T.Shibata "A 50nsec 64Kbit CMOS EEPROM with 200nsec

programming", in 5th Symp. VLSI Technol. Dig. Tech. Papers., p.76, May,1985.

(35)S.K.Lai, Y.W.Hu, S.Tam, G.K.Lum, and V.K.Dham, "Design of an EEPROM memory cell less than 100 square microns using 1-micron technology", in IEDM Dig. Tech. Papers, p.468, 1984.

(36)C.Kuo, K.Fu, P.Kim, M.Chonko, J.Jorvig, J.Yeargain, and J.Barnes "High Density FETMOS EEPROM cell using ONO inter-polysilicon dielectrics", in 5th Symp. VLSI Technol. Dig. Tech. Papers, p.76, May 1985.

(37)K.Shinada, Y.Nagakubo, K.Yoshikawa, and K.Kanzaki,"Current leakage through thermal oxide films grown on pattered poly-silicon layers", ECS Extended Abstruct, vol83-2, 1984,p.354.

(38)R.Cuppens, C.Hartgring, J.Verway, and H.Peek"An EEPROM for microprocessors and custom logic", in ISSCC Dig. Tech. Papers p.268, Feb.1984.

(39)S.Atsumi, S.Tanaka, K.Shinada, K.Yoshikawa, K.Makita, Y.Nagakubo, and K.Kanzaki, "Fast Programmable 256K read only memory with on-chip test circuits", IEEE J. Solid-State Circuits, vol.sc-20, no.1, p.422, Feb. 1985.

(40)N.Matsukawa, H.Nozawa, J.Matsunaga, and S.Kohyama, "Selective polisilicon oxidation technology for VLSI isolation", IEEE Trans. Electron Devices, vol.ED-29, no.4, p.561, Apr. 1982

(41)B.J.Hosticka, R.W.Broderson, and P.R.Gray, "MOS sampled data recursive filters using switched capacitor integrators", IEEE J. Solid State Circuits, vol.SC-12, p.600, Dec.1977.

(42)R.Gregorian, Y.A.Haque, R.Mao, R.Blasco, and W.E.Nicholson Jr., "CMOS switched capacitor filter for a two-chip PCM voice

coder", in ISSCC Dig. Tech. Papers, p.28, Feb.,1979.

(43)T.Sakurai, M.Isobe, T.Ohtani, K.Sawada, A.Aono, H.Nozawa, T.Iizuka, and S.Kohyama, "A Low power 46ns 256kbit CMOS static RAM with dynamic double wordline", IEEE J. Solid-State Circuits, vol. SC-19, no.5, p.578, Oct. 1984.

(44)E.Eichelberger and T.Williams. "A Logic Design Structure for LSI Testability", in Proceedings of the 14th Design Automation Conference, p.462, June 1977.

(45)J.Neal, et al."A 1Mb CMOS DRAM with Design for Test Functions" in ISSCC Dig. Tech. Papers, p.264, Feb., 1986.

(46)J.Beyers, et al."A 32-bit VLSI CPU Chip", IEEE J. Solid-State Circuits, vol.SC-16, no.5, p.537, Oct. 1981.

(47)F.Tsui, "The Cost and Speed Barriers LSI/VLSI Testing-Can they be Overcome by Testability Design", in 1985 International Test Conference, p.892, 1985

(48)Christopher Terman, "Simulation Tools for Digital LSI Design", Technical Report MIT/LCS/TR-304, MIT, Sept., 1983.

(49)J.Ousterhout, G.Hamachi, R.Mayo, W.Scott, and G.Taylor, "A Collection of Papers on Magic", Rept. UCB/CSD83/154, University of California at Berkeley, Berkeley, CA., Dec. 1983.

(50)C.Mead and L.Conway, "Introduction to VLSI Systems", Reading, Mass.: Addison-Wesley, 1980.

(51)M.Horowitz, P.Chow, D.Stark, R.Shimoni, A.Salz, S.Przybylski, J.Henessy, G.Gulak, A.Agarwal, and J.Acken, "MIPS-X: A 20-MIPS Peak, 32-bit Microprocessor with On-Chip Cache",IEEE J. Solid-State Circuits, vol.SC-22, no.5, p.790, Oct, 1987.

List of Figures

Fig. 1.1 Memory bit cost. 4

Fig. 1.2 Memory cell area as a function of minimum design rule. 4

Fig. 2.1 (a) Cell layout, (b) Equivalent circuit, and (c) Cross section view of ECL memory. 14

Fig. 2.2 Annealing time dependence of sheet resistivity. 16

Fig. 2.3 As^+ dose dependence on sheet resistivity. 16

Fig. 2.4 Annealing time dependence on lateral diffusion, y_j 16

Fig. 2.5 Memory block diagram of 256bit ECL RAM. 20

Fig. 2.6 Memory cell peripheral circuitry. 20

Fig. 2.7 Time dependent signal and bias voltage of internal nodes. 23

Fig. 2.8 Read current dependence of access time. 23

Fig. 2.9 Microphotograph of 256bit ECL RAM. 25

Fig. 2.10 Waveforms of address input and data output. 25

Tab. 2.1 Characteristics of 256bit ECL RAM. 26

Fig. 3.1 Schematic cross section of the N-well CMOS-bipolar transistors. 34

Fig. 3.2 DC characteristics of the NPN transistor. 34

Fig. 3.3 Cut-off frequency of the NPN transistor as a function of the collector current. 34

Fig. 3.4 Schematic diagram of CMOS RAM. 38

Fig. 3.5 Simulation results of the delay time in bitline and the sense amplifier as a function of bitline swing. 38

Fig. 3.6 Schematic of the bipolar and CMOS differential amplifiers. 40

Fig. 3.7 DC characteristics of the (a) bipolar and (b) CMOS differential amplifiers. 40

Fig. 3.8 Microphotographs of (a) bipolar and (b) CMOS differential amplifiers. 44

Fig. 3.9 Memory block diagram. 46

Fig. 3.10 Cell and its peripheral circuitry. 46

Fig. 3.11 Simulated wave forms of the 64K SRAM. Labels correspond to the circuit's nodes in Fig. 3.10. 48

Fig. 3.12 Voltage levels of the bit-lines and sense lines as a function of the supply voltage. 48

Fig. 3.13 Characteristics of the voltage regulator. 51

Fig. 3.14 Microphotograph of the (a) whole chip and (b) cells. 51

Fig. 3.15 Oscillographs of row and column access time. 53

Fig. 3.16 Current gain dependence on the collector current with and without the buried layer. 56

Fig. 3.17 Cut-off frequency dependence on the collector current, with and without the buried layer 56

Fig. 3.18 Delay time per stage of the ring oscillators. (a) CMOS 51 stage ring. (b) ECL 18 stage ring. 58

Fig. 3.19 Bi-CMOS buffer circuit. 58

Tab. 3.1 Speed comparison of the bipolar and CMOS amplifiers. 44

Tab. 3.2 Characteristics of high speed CMOS 64k SRAM. 53

Fig. 4.1 Layout patterns of new single polysilicon EEPROM cells. (a) Cell A, and (b) Cell B. 70

Fig. 4.2 Schematic cross section view of the cells. (a) a-a' of the cell A, and (b) b-b' of the cell B. 70

Fig. 4.3 Simulation results of the threshold shift in the erase operation, in case of neglecting the tunneling effect through the diffused control gate (thin line), or including the effect (fat line). $d_1 = d_2 = 83\mathring{A}$. 75

Fig. 4.4 Effective voltages applied to the thin oxide as a function of the impurity concentration, N_D, below the thin oxide,when the depletion region is built. 75

Fig. 4.5 Waveforms of the floating gate, threshold shift, and the tunneling current in both the erase and program operation. 79

Fig. 4.6 Threshold voltage shifts and peak tunneling current as a function of the rise time of the control gate, and the drain voltages. $d_1=d_2=83\mathring{A}$, $V_C=V_D=15V$, and $t_e=t_P=2ms$, excluding the rise and fall time. 79

Fig. 4.7 Erase/Program characteristics of cell A as a function of the applied voltage. 81

Fig. 4.8 Erase/Program characteristics of Cell B as a function of the erase and program time. 81

Fig. 4.9 Gate and substrate current of the thin oxide transistor. The drain profile was fabricated similar to the floating gate transistor of cell B. 84

Fig. 4.10 Endurance characteristics of cell A (broken line) and cell B (solid line). 84

Fig. 4.11 Layout pattern of a new single-polysilicon EEPROM cell, named DIFLOX. The cell is composed of a selected gate and a floating gate transistors. 87

Fig. 4.12 Cell array of DIFLOX. The selected gate is equivalent to the word-line, and the control gate is connected to the program-line (PL) via a pass transistor. 87

Fig. 4.13 Cross section views of DIFLOX, along a-a', b-b', and c-c', in Fig. 4.11. 89

Fig. 4.14 Turn-on voltage of the field transistors as a function of the distance. It is defined as the leakage current exceeding 10^{-10} A/um. The drain is biased to 20V. 89

Fig. 4.15 Threshold shift of DIFLOX in the erase/program state as a function of a word-line voltage. V_{PP} = 15V, d_1 = 85Å. 91

Fig. 4.16 Endurance characteristics of DIFLOX. V_{PP} = 15V, d_1 = 85Å, V_{WL} = 20V. 91

Fig. 4.17 Microphotograph of DIFLOX cell array. The cell size is 7.5*11.5um. 93

Fig. 4.18 Block diagram of 256kbit EEPROM in the read operation. 94

Fig. 4.19 Cell peripheral circuitry. Improved open-bit-line scheme and distributed sense amplifiers are shown. 94

Fig. 4.20 Monitored waveforms of internal nodes, obtained by the electron beam tester. s_0 and s_1 are two nodes of a sense amplifier, and D_0^* corresponds to the output signal before the final buffer. The levels are incorrect, not only absolutely,but relatively, because they are influenced by the electron scattering of the neighbor patterns in the LSI. 97

Fig. 4.21 Cell peripheral ciruitry in erase/program operation. 100

Fig. 4.22 Timing chart of the page-mode programming and the data polling. 100

Fig. 4.23 Successive data loading control circuits (Φ_{START} generator). A period of 100us is counted by the switched capacitor. 102

Fig. 4.24 (a) Timer circuit, and (b) wave forms at several nodes for erase/program. Improved switched capacitor is used. It counts 1.8ms lapse time. 102

Fig. 4.25 Output characteritics of high-voltage generator. In the memory, the basic frequency by the internal oscillaor is set to 10MHz. 105

Fig. 4.26 Output characteristics of high-voltage pump, It is attached to every bit and word line. 105

Fig. 4.27 Microphotograph of the 256kbit EEPROM whole die. The die size is 7.33*6.23mm. 107

Fig. 4.28 Monitored waveforms of Φ_{STA} and V_{PP}. V_{PP} is raised to 20V. 107

Fig. 4.29 Oscillograph of the address and data-output waveforms with external loads. 109

Tab. 4.1 Comparison of device performance between cell A, and cell B. 85

Tab. 4.2 Characteristics of 5V only 256kbit EEPROM. 109

Fig. 5.1 A simple tester built using DGR chips. 117

Fig. 5.2 A block diagram of the DGR. 117

Fig. 5.3 Schematic of the memory showing peripheral circuits. 123

Fig. 5.4 Timing waveforms for the memory during synchronous operation. 123

Fig. 5.5 Typical measurement sequence showing the parallel memory access and the parallel to serial conversion. 125

Fig. 5.6 Timing waveforms for Jump and Branch operation. 125

Fig. 5.7 Bit-line swing as a function of bit-line current. 128

Fig. 5.8 Short circuit protection. 128

Fig. 5.9 Schemation of the adjustable delay circuit. 130

Fig. 5.10 Elemental adder circuits. 133

Fig. 5.11 Address generator. 133

Fig. 5.12 CAD tools. 135

Fig. 5.13 Die photo of 3um DGR chip. 138

Fig. 5.14 Address access time as a function of supply voltage. The delay are given both for slow (D0) and fast (D8) data lines and full memory access (A3) and column select (A0). 138

Fig. 5.15 Maximum vector rate as a function of supply voltage. 140

Fig. 5.16 DGR output waveforms of several working modes. Φ_1 and Φ_2 are the inputs. It is shown that DGR is working over 10MHz vector rate in every mode.
(a) Φ_1 and Φ_2 synchronous waveforms
(b) RZ and NRZ waveforms,
(c) Jump flag (JMP). 140

Fig. 5.17 VOH/VOL versus supply/sink current of DUT pin. 141

Tab. 5.1 Test sequences. 114

Tab. 5.2 DUT control registers (100-11F).
A-reg.:even B-reg.:odd. 120

Tab. 5.3 Address registers. 120

Tab. 5.4 Status registers. 120

Tab. 5.5 Design rules. 135

Tab. 5.6 Performance of 2um and 3um DGR. 141

Publications

(1)H.Abe, J.Miyamoto, and R.Itatani, "Grid Effects on the Plasma Simulation by the Finite-sized Particle", J.Comp.Phys, vol.19 No.2, p134, Oct.,1975.

(2)H.Ikoma, K.Ichinose, K.Kanzaki, S.Shinozaki, K.Fujita, J.Miyamoto, Y.Miyasako, "Introduction of VLSI (4), Bipolar Integrated Circuits", 1984, Kindai-Kagaku-sha. (in Japanese)

(3)S.Shinozaki, T.Iizuka, F.Masuoka, K.Shinada, and J.Miyamoto,"Role of the External n-p-n Base Region on the Switching Speed of Integrated Injection Logic(IIL)", IEEE J. Solid-State Circuits, vol.SC-12, no.2,p.185,April 1977.

(4)S.Shinozaki, K.Shinada, and J.Miyamoto, "Effects of Gate Geometry on Propagation Delay of Integrated Injection Logic (IIL)", IEEE J. Solid-State Circuits, vol.SC-13, no.2, p225, April 1978.

(5)T.Sugano, T.Sakurai, H.Ohiwa, M.Sugawara, K.Natori, A.Kanuma, R.Dan, J.Miyamoto, S.Saitoh, "Introduction to MOS LSI Design", (translated in Japanese), 1983, Sangyou-Tosho.

(6)J.Miyamoto, "Recent Development and Application of 64K and 256K EEPROM", lecture text, Japan Technology Information Center, 1985. (in Japanese)

(7)J.Miyamoto, K.Shinada, S.Shinozaki, and N.Sekiguchi, "Application of Polycrystalline Silicon Load For High Performance Bipolar Memory", in IEDM (International Electron Devices Meeting) Tech., Dig., p50, Dec. 1980.

(8)J.Miyamoto, S.Saito, H.Momose, H.Shibata, K.Kanzaki, and S.Kohyama, "1.0 um N-Well CMOS/Bipolar Technology for VLSI Circuits", in IEDM Tech.,Dig., p.63, Dec. 1983.

(9)J.Miyamoto, S.Saito, H.Momose, H.Shibata, K.Kanzaki, and T.Iizuka, "A 28ns CMOS SRAM with Bipolar Sense Amplifiers", in International Solid-State Circuit Conference(ISSCC) Dig. Tech. Papers, p.224, Feb. 1984.

(10)J.Miyamoto, S.Saito, H.Momose, H.Shibata, K.Kanzaki, and T.Iizuka, "A High Speed 64K CMOS SRAM with Bipolar Sense Amplifiers" ,IEEE J. Solid-State Cicuits, vol.SC-19, no.5, p.557, Oct.,1984

(11)H.Momose, H.Shibata, S.Saito, J.Miyamoto, K.Kanzaki and S.Kohyama, "1.0 um N-well CMOS/Bipolar technology", IEEE, J. Solid-State Circuits, vol. SC-20, no.1, p.137, Feb. 1985.

(12)N.Matsukawa, S.Morita, K.Shinada, J.Miyamoto, J.Tsujimoto, T.Iizuka,and H.Nozawa, "A High Density Single Polysilicon Structure EEPROM with LB(Lower Barrier Hight) Oxide for VLSI's", in 5th,Simp. VLSI Technol. Dig., Tech. Papers, p100, May 1985.

(13)J.Tsujimoto, J.Miyamoto, N.Matsukawa, K.Shinada, S.Morita, H.Nozawa,and T.Iizuka, "A 5V-Only 256K CMOS EEPROM using Barrier Height Lowering Technique",in 11th European Solid-State Circuits Conf. Dig., Tech., p.241,Sept.1985.

(14)J.Miyamoto J.Tsujimoto, N.Matsukawa, K.Shinada, S.Morita,H.Nozawa,and T.Iizuka, "An Experimental 5-V-Only 256-kbit CMOS EEPROM with a High-Performance Single-Polysilicon Cell",IEEE J. Solid-State Circuits, vol.SC-21, no.5, p.852, Oct.,1986.

(15)J.Miyamoto and M.A.Horowitz,"A Single-Chip Functional Tester", in ISSCC Dig. Tech. Papers, p.232, Feb., 1987.

(16)J.Miyamoto and M.A.Horowitz, "A Single-Chip LSI High-Speed Functional Tester", IEEE J. Solid-State Circuits, vol.SC-22, no.5, p.820, Oct., 1987.

(17)N.Ohtsuka, S.Tanaka, J.Miyamoto, S.Saito, S.Atsumi, K.Imamiya, and T.Iizuka, "An OTP test circuit on the 4Mb CMOS EPROM", in Symp.VLSI Circuit Dig. Tech. Papers, 1987,p55

(18)N.Ohtsuka, S.Tanaka, J.Miyamoto, S.Saito, S.Atsumi, K.Imamiya, K.Yoshikawa, N.Matsukawa, S.Mori, N.Arai, T.Shinagawa, Y.Kaneko, J.Matsunaga, and T.Iizuka, "A 4-Mbit CMOS EPROM", IEEE J. Solid-State Circuits, vol.SC-22 p.669, Oct.1987.